Replication in Research on Quality in Conference Interpreting

Franz Pöchhacker
University of Vienna

With reference to survey research on quality in conference interpreting, which constitutes a cohesive line of investigation in interpreting studies, the paper highlights the role of replication in the scientific process and reviews examples of studies carried out since the 1980s. With the pioneering survey by Bühler among members of the International Association of Conference Interpreters serving as the starting point, a number of studies among end-users of conference interpreting are mentioned before the focus of the paper shifts to research efforts aimed at replicating Bühler's work on quality criteria as seen from the perspective of professional interpreters. In the analysis, which places the emphasis on methodological issues but also presents some relevant findings, several replication studies are closely examined, and various shortcomings as well as advances in research design are discussed.

Keywords: quality criteria, survey research, duplication, methodology, AIIC

1. Introduction

In the development of research on interpreting, which received a strong impetus from specialists in other disciplines particularly in its early stages, scientific aspirations for the field have been closely associated with *innovation* through empirical research. Indeed, scientific research, by definition, is expected to produce *new* findings and insights, so that the value and vitality of a discipline may be judged by the extent to which it generates new discoveries and fresh knowledge. However, while the role of innovation as a supreme value in the scientific endeavor is hardly in doubt, it cannot usefully be regarded in

isolation. Rather, it must be linked to other key principles of scientific research, first and foremost among them being validity. The validity of new findings, as achieved, for example, in an experimental study or through a survey, may be undermined in many ways, by design flaws and confounding variables, so one important strategy for ascertaining their reliability is to repeat the study, which can broadly be referred to as *replication.*

The notion of replication – and of reliability and validity, for that matter – can be rather complex, and it is the aim of this paper to discuss the role of replication in research design and consider various forms of its implementation. Not being able to claim any special methodological expertise, I will be approaching this topic from an applied perspective. Rather than methodology per se, my focus will be on examples of actual empirical research − in this case of research on quality in conference interpreting. Within the field of interpreting studies, this focus is admittedly narrow, in several ways: (1) it is limited to international conference interpreting; (2) it concentrates on quantitative survey research; and, more fundamentally, (3) it adopts a "classic" perspective on empirical science, as represented, for instance, by Popper's critical rationalism. To a large extent, these choices are shaped by the topic as such, since the idea of replicability is a core assumption of empirical science in the first place, and becomes increasingly questionable, or even antithetical, the more one embraces the tenets of qualitative research (e.g. Denzin and Lincoln 2000), where the focus is on the individual case and the subjective construction of meaning. As I have argued elsewhere (Pöchhacker 2011), interpreting scholars, in a postmodern spirit, need not view the quantitative and qualitative paradigms of science as mutually exclusive and can adopt a pragmatic stance that allows them to choose the approach that best suits the topic and problem at hand.

2. Replication as research

2.1. Replication in research design

The notion of replication is sometimes associated with experimental research, or with studies investigating causal relationships, but can be defined more generally as "a duplication of a previously published empirical study" (Hubbard

and Armstrong 1994: 236). Whereas the term "duplication" might suggest a lack of value, as in "a mere copy" of something or a superfluous additional effort, the role of replication in scientific research is exactly the opposite − that is, a way of ensuring that the original study has value in the first place. Rather than "value," the notion of *validity* is central to these considerations, and replication fundamentally aims to establish that the initial results are valid by showing that they can be reproduced.

The term "repetition" is loosely used as a synonym, and La Sorte (1972: 218) initially defined replication as a "conscious and systematic repeat of an original study." Most authors nowadays make a distinction between internal repetition, i.e. retesting within a given study, and replication proper, implying that genuine replication should be done independently, by different researchers in a different environment. After all, replication is considered a key instrument in combatting scientific fraud as well as a tool for discovering bias and errors in research design. Not without reason, therefore, replication is a hallowed principle of scientific research. And yet, it is a principle that is not very often practiced.

The reasons for the lack of replication studies − in most disciplines − are manifold. To begin with, they are not easy to do, considering the need to replicate all relevant facets of the original work. Depending on the object of study, access to materials, including measurement instruments, and human subjects may be limited, even in the ideal case that the original research report was sufficiently detailed and explicit. And if and when they have been successfully accomplished, replication studies in many fields are less likely to be published by the most prestigious journals. As a result of such editorial practices − and of novelty-seeking attitudes in general, doing replications is of limited appeal within the scientific community, regardless of their outcome. When the replication corroborates the initial findings, it is likely to add to the credit of the original authors; in the opposite case, the upshot in terms of academic relations may also be problematic.

Despite these difficulties, which may explain why replication studies are not very common, the intrinsic value of repeating previously published work to corroborate existing findings is beyond doubt. A simple example of such direct replication duplicating all facets of the original study would be a survey in which the same instrument is administered to another sample drawn from the same population. More often than not, however, some facets of the original study conducted in a given place at a certain time are difficult to reproduce

by researchers in a different environment. Therefore, whether by choice or necessity, some facets of the study may differ in the replication, and its purpose will no longer be to confirm the internal validity of the original work but to extend its external validity, or generalizability (Hubbard and Armstrong 1994: 236).

For an object of study like interpreting, which is highly specialized and characterized by tremendous variability in terms of languages, settings, topics, genres and individual expertise, strict replication of an experimental study or even a survey is an enormous challenge, so one would typically expect what Morrison et al. (2010: 282) label "partial replications." These involve the modification of some aspect or variable of the original research design and therefore serve to investigate the extent to which the initial findings will hold under different circumstances. Hubbard and Armstrong (1994: 236) refer to this as "replication with extension." As explained by Hubbard and Vetter (1997: 3),

> *[t]he major goal of extensions is to assess whether earlier results are capable of being generalized to other populations, product categories, time periods, organizations, measurement instruments, geographical areas, investigators, and so on, as opposed to being idiosyncratic or localized in nature.*

A number of further distinctions can be made to characterize different types and grades of replication. These include "conceptual replication," in which the researcher aims to confirm the original findings using a different methodological approach, and "replication with update." The latter is described by Morrison et al. (2010: 282) as a replication in which certain aspects of the initial empirical study are modified in line with changes in the research environment, as in the case of a standardized test that can be used several decades later only with appropriately modified material references and language. As the authors point out, a replication with update, unlike a partial replication, supersedes rather than complements the original study, as it is deemed more valid in the current environment.

2.2. Replication in interpreting research

As indicated above, there is a striking discrepancy between the high regard for replication as a crucial factor in the scientific process and the low rate at which replication studies are actually carried out. The lack of replication has been lamented for many different fields, from business studies and sociology to psychology and nursing research. Interpreting studies is no exception, and Daniel Gile (1998) observes that, typically, in interpreting studies "investigators are interested in original research, but much less in replication." In a more metaphorical vein, Dodds and Katan (1997: 90-91) characterized the lack of scientific verification of proposals for teaching and testing as follows:

> *What seems to be happening is that having reached the Moon, nobody is any longer that interested in it. Its exploration and colonisation are of secondary importance, forms of life there have become irrelevant because it seems we must proceed at all costs with great leaps and bounds.*

The simile used here aptly captures the concern with discoveries and new findings, as these seem to be a much better reflection of progress and scientific advances than the footwork involved in replication for the purpose of confirming and consolidating initial findings.

The importance of replication has also been stressed by Gile, with particular regard to experimental research (Gile 2005), but also pointing to the role of replication in providing the research community with "a means of assessing the representativeness of data obtained in the single studies" (Gile 1990: 230). In line with this understanding of replication, which would involve some type of extension so as to test the generalizability of available findings, Ingrid Kurz (2001) describes "repeating/modifying a previous study" as a useful approach for designing a research project and gives the example of a follow-up study on the topic of videoconference interpreting. Interestingly, Kurz chooses a different heading to introduce − and illustrate with an example of her own work − yet another approach to developing a study. In a section entitled "reexamining other people's conclusions" she recalls the genesis of her user expectation surveys in reaction to the work of Hildegund Bühler (1986), who had concluded that the quality-related preferences expressed by members of the International Association of Conference Interpreters (AIIC) would also

reflect the requirements of the users. Her doubts that the interpreters' and the end-users' perspectives on quality would be the same prompted her to use the first part of Bühler's questionnaire in a survey among participants at a medical conference. While the equal number of respondents in the two small-scale surveys appears to have been coincidental, Kurz's (1989) use of eight identical questionnaire items together with the original four-point ordinal rating scale in a different population could well be regarded as a case of replication with extension. Indeed, her first user expectation survey may be regarded as the crucial (partial) replication study in survey research on quality criteria for conference interpreting, which is the main topic of this paper to which I now turn.

3. Replication in survey research on quality criteria

3.1. From interpreters to users

The user expectation study conducted by Kurz (1989) on the basis of Bühler's (1986) survey on quality criteria among AIIC members extended the target population from conference interpreters to conference interpreting users and could thus be classified as a partial replication in the technical sense introduced above. While Kurz used the same criteria and rating task, she did not use all the 16 items in Bühler's original instrument. In this respect, her replication study could also be qualified as partial in the more common sense of the term. Moreover, by omitting the items relating to professional behavior and administering the survey in a conference with simultaneous interpreting services, Kurz narrowed the focus to output quality in the simultaneous mode. Bühler's study, in contrast, sought to establish the criteria underlying AIIC members' peer evaluation of candidates for membership, whose skill profile would necessarily encompass consecutive and simultaneous interpreting. These modifications ought to be taken into account when taking a comparative look at the findings, shown for the output-related quality criteria in Figures 1 and 2.

While a detailed discussion is beyond the purpose and scope of this paper, it is easy to see that the ratings given by the interpreters are consistently higher. Only the top three criteria, which hold identical ranks in both studies, are considered "(highly) important" by at least 80% of the medical-conference

participants, whereas all but one criterion (native accent) command such high ratings in the survey among AIIC members. Given the above-mentioned modifications in the study design, the higher expectations among the professionals may also have to do with ideal standards in regulating access to the profession as opposed to expectations that may have been shaped by users' overall satisfaction with the interpreting services received.

Figure 1. Quality criteria as rated by 47 AIIC members (based on Bühler 1986)

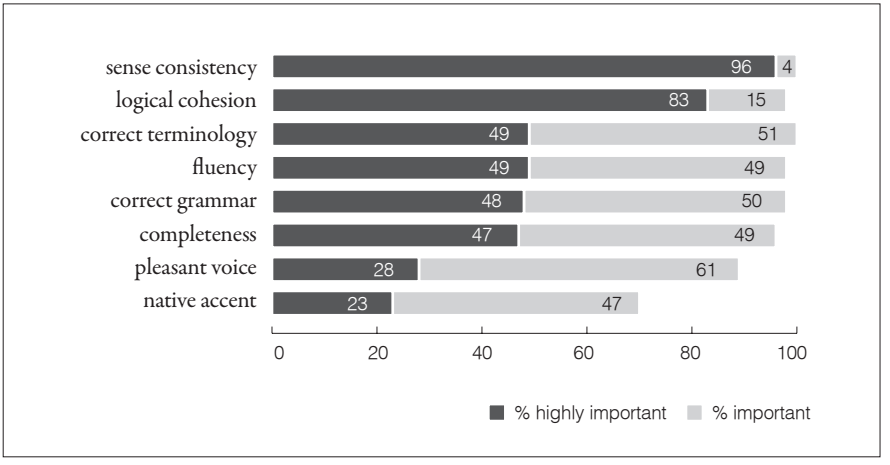
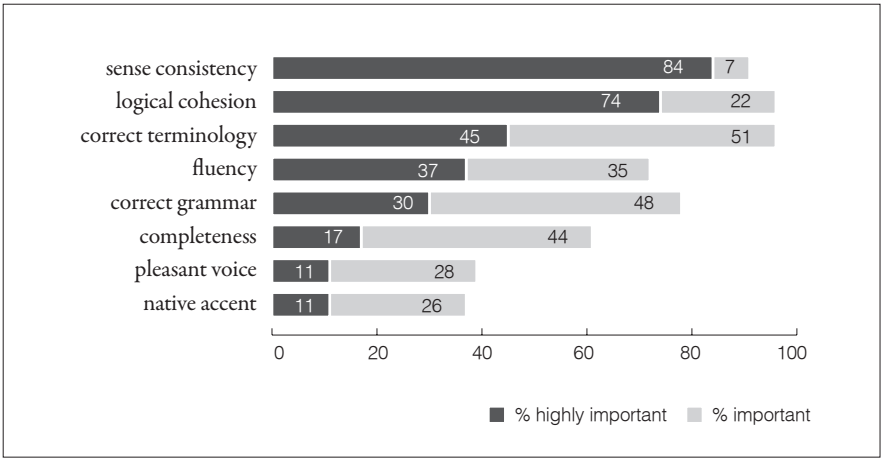


Figure 2. Quality criteria as rated by 47 conference participants (based on Kurz 1989)



Another hidden variable in this comparative analysis is language. The fact that Kurz also used a German version of the questionnaire is a consequence of her extension effort, as some members of the target population may not have been fully comfortable with English as the language of the questionnaire. This translational component, however, is by no means trivial, as should be evident particularly within the community of translation scholars.

Even within a given language, the wording used to refer to a certain (or, as we shall see, uncertain) construct may pose major challenges to the validity of a study. This issue is acknowledged by Bühler (1986: 232) in her original article, as she had reason to suspect from some of her findings that terms like "completeness" may have been open to (mis)interpretation, making it quite uncertain what constructs the respective questionnaire items were actually measuring. Her doubts were echoed in a comment by Danica Seleskovitch (1986), who also questioned the notions of "correct grammatical usage" and "correct terminology," and doubts regarding the latter were raised in particular by Mack and Cattaruzza (1995). In light of these conceptual uncertainties, translation is likely to bring a further complication. In a reanalysis of Kurz's (1989) data by language groups it was found, for instance, that the criterion of completeness received significantly lower ratings from respondents using the German version of the questionnaire (Pöchhacker 2005: 156). Whether this could have been due to the syntactic and lexical shifts in the German translation, in which "completeness of interpretation" was rendered more redundantly as "complete rendition of the original" *(vollständige Wiedergabe des Originals),* is hard to establish. At any rate, this example serves to heighten our awareness of the role of translation in replication studies. Whereas many social scientists may still regard this merely as a code change with little effect on their measurement instrument, the issue of translation in multilingual survey research, which is often a case of replication by extension across linguistic and – cultural – boundaries, clearly needs to be given due attention in our present context. Indeed, transposing a survey to a different socio–cultural environment might even be considered analogous to a replication with update. Where extensive adaptation to the target-cultural environment is required, the replication study might no longer be assumed to merely complement the original study. This may suggest the need for a special type of replication that one could refer to as "replication with cultural adaptation." A meta-analysis of such studies would then need to account for the unique features of the cultural

environment rather than simply aggregating the findings for the different populations.

3.2. More on users

In addition to the idea of comparing the ratings of interpreters and end-users, Kurz also sought to ascertain whether different groups of users differed in their quality-related preferences. She did so by distributing her questionnaires also at an engineering conference and at a conference on education under the auspices of the Council of Europe. The results of these studies, which suggest some variability in the patterns established for different user groups, have been widely discussed and need not be reiterated here (see Kurz 1993). What has not been discussed is whether these studies come under the heading of replication. More so than Kurz's extension of Bühler's study, the second and third surveys could be labeled as replications with extension, or partial replications (using a different population), as far as the design as such is concerned. On the other hand, the fact that all surveys were conducted by the same individual researcher, and presented in a single report focusing on between-group differences, makes this a case of different measurements within a single study design. While this would be accepted as a "retest replication" in the scheme proposed by La Sorte (1972), such studies are no longer regarded as genuine replications (cf. Thompson 1994), quite apart from the fact that the requirement of *independent* duplication is obviously not met.

At any rate, the study by Kurz (1993) helped "initiate a whole new line of investigation" (Kurz 2001) which includes examples of conceptual replication employing different methodological techniques (e.g. Vuorikoski 1993; Kopczyński 1994; Moser 1996) but relatively few cases of direct replication. One attempt at strict replication was undertaken by Gabriele Mack and Lorella Cattaruzza (1995), who adopted the multimethod approach used by Anna-Ritta Vuorikoski (1993) in studying the on-site experience and general expectations of simultaneous interpreting among participants in interpreter-mediated seminars in Finland.

Another particularly interesting case of replication in research on quality expectations among users is the study by Ángela Collados Aís (1998), who surveyed 42 legal scholars in Spain as part of a more complex experiment. Her questionnaire included the items used by Kurz (1993) but also additional

ones, such as "monotonous intonation," which constituted the focus of the experimental study. Several modifications to the original study design are worth noting: (1) the questionnaire was offered in a Spanish version; (2) the items were formulated in the negative sense ("lack of x"), in line with the (negative) criterion of interest ("monotonous intonation"), so that the question focused not on the *importance* of a given criterion but on the degree of its *negative impact;* and (3) ratings were requested on a five-point scale, from "1" (= most) to "5" (least). Aside from these modifications, the target population of the study – Spanish academics in the field of law – clearly make this study a replication with extension, the results of which, incidentally, largely confirmed the pattern of user preferences found in previous studies.

The research efforts led by Collados Aís are especially noteworthy also because they involved several types of internal replication – or extension: For one, the questionnaire described above was also administered to a group of 15 professional interpreters and interpreter trainers at the University of Granada, thus reestablishing the contrastive view of user and interpreter perspectives as in the initial study by Kurz (1989). What is more, the entire expectation vs. assessment study was replicated with another group of users, again with a legal background, by a research team involving colleagues at other (Spanish) universities (Collados et al. 2007). As such, the project led by Collados Aís constitutes an outstanding example of replication in research on quality in interpreting, and in the field of interpreting studies in general. Its expectation survey component once again confirmed the sequence of importance established in similar studies; that is, "sense consistency" (or "correct rendition," in the Spanish version) and "logical cohesion" at the top, followed by fluency, correct terminology and completeness on a second tier, and paraverbal aspects such as intonation, pleasant voice and native accent as the least important criteria, ranking below such aspects as correct grammar, diction and appropriate style. Thanks to these studies, the output-related criteria of performance quality proposed by Bühler (1986) have indeed been investigated in numerous studies among users of simultaneous interpreting, and replication has yielded its fundamental benefit of confirming original findings (e.g. Kurz 1993) and extending them to a larger population.

In the very same study, however, replication also served to invalidate a previous finding. Whereas Collados Aís (1998) had found that poor intonation had a significant negative impact also on users' assessment of overall

performance quality, no such effect was found in the replication study (Collados Aís 2007). Here again, though, the value of replication in research is obvious, and it is to the credit of the researcher in question to have undertaken such efforts and corrected, or qualified, a finding for which her original study had been widely cited.

From this most impressive example of replication in research on quality in interpreting, in this case among users, I would like to turn back to the service provider perspective and review three cases of replication research based on the seminal study by Bühler (1986). These are the web-based survey by Chiaro and Nocella (2004), discussed critically in a previous paper (Pöchhacker 2005); a recently completed MA thesis conducted in France (Jolibois 2010); and a comprehensive survey study carried out at the University of Vienna in the context of a larger research project on "Quality in Simultaneous Interpreting" (QuaSI 2010).

4. Bühler revisited

4.1. Rating vs. ranking

The survey on quality criteria published by Delia Chiaro and Giuseppe Nocella (2004) was conducted in late 2000 as a pioneering effort in online research within the field of interpreting studies. The authors designed a web-based questionnaire and sent some 1,000 e-mail invitations containing the link to the questionnaire "to interpreters belonging to several professional associations" (2004: 284). Chiaro and Nocella indicate that they used "several spammings" and note that "[t]he e-mail addresses of these interpreters were gathered visiting the websites of the interpreters" (2004: 292). Even so, it remains regrettably unclear just how the target population of their survey was defined. The professional associations in question remain unidentified, so it is not clear whether or to what extent AIIC members were included in the sample. This is highly problematic from the perspective of replication, as it is not possible to establish whether the authors conducted a direct replication in the same population or extended the study to a different one, albeit insufficiently defined. There is some evidence pointing to the latter, including the fact that as many as one fifth of the e-mail invitations resulted in delivery

failures, and the finding that, apparently, "most respondents do not interpret into their mother tongue" (2004: 285).

Aside from the choice of target population, the survey by Chiaro and Nocella (2004) fails to meet the key requirements of a direct replication also in other respects. Most importantly, and as a matter of explicit choice, the authors replaced Bühler's (1986) rating task with a force-choice ranking so as to elicit a clear order of priorities. This modification of the study design is of great potential value, provided that the replication – with methodological extension – keeps other relevant variables unchanged. A partial replication of this sort would thus serve to examine whether the original findings can be validated with a different set of response options in the questionnaire. But given the problem of the ill-defined population, the study does not reliably lend itself to such a comparison. What is more, Chiaro and Nocella (2004) also used a different set of criteria in their questionnaire, notwithstanding their claim that "[t]he criteria used in this investigation are the same as those used by Bühler" (2004: 283). The modifications they mention in continuation relate to their use of a rank order scale and the fact that the ranking task required them to offer Bühler's "linguistic" and "extra-linguistic" criteria in two separate sets. In fact, though, most of the original extra-linguistic criteria were replaced by different ones, and Chiaro and Nocella also reformulated some of the output-related ("linguistic") criteria, such as "completeness of interpretation," which becomes "completeness of information," and "sense consistency with the original," shortened to "consistency with the original."

Table 1. Comparative Ranking of Quality Criteria

Chiaro & Nocella (2004)	**Bühler (1986)**
1. consistency with the original	sense consistency with original message
2. completeness of information	logical cohesion of utterance
3. logical cohesion	use of correct terminology
4. fluency of delivery	fluency of delivery
5. correct grammatical usage	correct grammatical usage
6. correct terminology	completeness of interpretation
7. appropriate style	pleasant voice
8. pleasant voice	native accent
9. native accent	appropriate style

The findings from that study, limited here to output-related aspects (as in Figures 1 and 2), are shown in Table 1 in the form of a list, juxtaposed with one compiled on the basis of Bühler's percentages for ratings of "(highly) important."

Without going into a detailed discussion of these findings, the difference regarding the relative importance of "completeness" (ranking second vs. sixth) and also of "correct terminology" (ranking sixth vs. third) is striking. The replication study has clearly been productive in yielding a new pattern of findings compared to Bühler's (1986) original survey among AIIC members. However, finding a meaningful explanation for these differences is made very difficult by the multiple modifications in the study design and by the uncertainties regarding the survey population. One might conjecture that the web-based survey also reached interpreters working in non-conference settings, including courts and police, for which standards of accuracy and completeness – and interpreting modes, for that matter – may differ from those typical of international conference interpreting. But without explicit and detailed information about all aspects of the study design, Chiaro and Nocella's (2004) "multiply partial" replication of Bühler's study does not allow us to reap the full benefit that replication could deliver.

4.2. "Duplication"

Yet another example of a study replicating Bühler's (1986) quality criteria survey among interpreters can be found in a Master's thesis completed at the University of Burgundy for a degree in "Languages and Business" (Jolibois 2010). In a project ostensibly centered on the issue of the interpreter's role, the author, who confesses to a "very partial knowledge of interpreting" (2010: 51), included a quality criteria rating task among a total of 16 questions, all of which were addressed to AIIC members in a full-population survey. By virtue of its clearly defined population, the survey by Simon Jolibois could count as a direct replication of Bühler's study, using online techniques to reach a larger group of respondents at no additional expense. Unfortunately, the young researcher introduced major modifications that largely erode the potential gains from this replication. Without much explanation, if any, he changed Bühler's list of criteria, dropping "logical cohesion" and "correct grammatical usage" and

adding such items as "imagination," "capacity of improvisation," "elegance," "charisma" and, most curiously, "smile." What is more, the importance of these "features of an interpreter" (Jolibois 2010: 81) were to be rated on a five-point scale ranging from "essential" and "important" on the positive side to "limited influence" and "irrelevant for an interpreter" on the opposite end, with a broad middle ground labeled "neutral (desirable but not fundamental for the quality of interpreting)."

While the researcher is of course free to refine the survey instrument as needed, it seems problematic to introduce modifications without appropriate justification. Since the criterion of logical cohesion, for instance, is among the top three in most surveys, the decision to do without it should have been properly explained. What is more, a close look at the rating scale shows a mix of conceptual dimensions, ranging from "importance" to "influence," and relating to "interpreting" in one case and to "an interpreter" in another. Indeed, it may be hard to clearly distinguish between a feature that is "desirable but not fundamental" and one that is of "limited influence." Most critically, however, the author treats what purports to be an ordinal scale as an interval scale and calculates arithmetic means instead of reporting the results as percentages – as done, incidentally, by Kurz (1993) and rightly challenged by Chiaro and Nocella (2004).

The author's error in the statistical treatment of the results could be remedied by a reanalysis of the data, and his use of modified response options could claim the merit of replicating Bühler's findings (for the criteria left unchanged) with a different measurement scale, as in the case of Chiaro and Nocella (2004). Nevertheless, this replication suffers from a more fundamental weakness, namely a very low response rate. While the absolute number of respondents – 189 – seems high in comparison with the sample size in Bühler's (1986) original study, it amounts to only 7.5% of the target population (i.e. AIIC members with English in their language combination). Unfortunately, the author did not ask for such basic background data as age, gender or working experience, so it is difficult to assess to what extent his sample is representative of the overall population. In terms of employment status, staff interpreters appear to be somewhat overrepresented (with 15.3% compared to the share of 10% reported by the Association (AIIC 2006).

The author's failure to elicit socio-professional background information could be excused with reference to Bühler's (1986) study, which included no

such data whatsoever. What is exceptionally awkward, however, is the fact that Jolibois conducted his AIIC survey on "the role of conference interpreters" twenty months after a similar, well-publicized effort by our research team at the University of Vienna (Zwischenberger et al. 2008), which will be presented in the following section. Contrary to what Jolibois asserts in his report, the survey administered by Zwischenberger was devoted to the interrelated issues of quality and role (comprising a partial replication of Angelelli 2004) rather than "simply quality" (Jolibois 2010: 66). The author thus duplicated the research effort undertaken at the University of Vienna unwittingly, with duplication in this case amounting to less than a replication, recalling that replication is defined as a "conscious and systematic repeat of an original study" (La Sorte 1972: 218). Such duplication, in the common, negative sense, could be criticized as an imprudent use of valuable academic resources. More consequentially, however, it may have been a factor contributing to the disappointingly low response rate (aside from others, such as the poor layout and user-unfriendliness of the questionnaire). As much as one appreciates the willingness of nearly 200 professionals to spend time answering yet another questionnaire, academics should be as respectful as possible of conference interpreters as partners in research, and use this resource with great care and efficiency. In the case under study, a basic standard of care, implying background research, communication and coordination, appears to have been missing, for it is hard to reconcile the author's assertion that he only found out about the earlier survey in July 2010 (Jolibois 2010: 66) with the fact that the report on that study was published in March 2010 on the AIIC website (Zwischenberger & Pöchhacker 2010), where Jolibois claims to have compiled his list of e-mail addresses.

Without wishing to come down all too harshly on a junior colleague who has undertaken an ambitious study (apparently supervised by a non-specialist in interpreting research), this case of a missed opportunity for effective replication research in the field of interpreting deserves exposure for the purpose of avoiding similar mistakes in the future.

4.3. Getting it right

The third and final example of a study replicating Bühler's seminal survey stems from a grant-funded project carried out at the University of Vienna

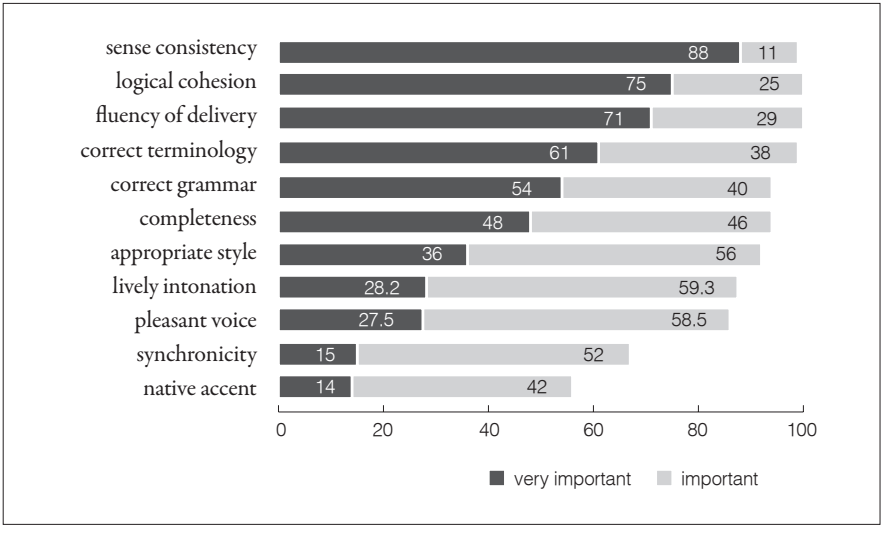
(QuaSI 2010). It is worth noting that the aim of replicating a previous study was stated explicitly in the grant application to Austria's top-level science funding agency. This suggests that the Science Fund, or rather, the two international peer reviewers it commissioned to assess the merits of the research proposal, regarded replication as a worthwhile scientific endeavor in its own right.

The goal of our study (which was part of larger project that also included innovative experiments) was both to replicate Bühler's survey and to go beyond it in several ways. With regard to the replication as such, our goal was simply to "get it right," duplicating Bühler's single-handed effort of a quarter-century ago in a manner that would live up to 21st century standards, if not set new ones. Mindful of the pioneering effort by Chiaro and Nocella (2004), we opted for a web-based survey but used state-of-the-art software to ensure controlled access to the questionnaire as well as anonymity of the responses. The LimeSurvey questionnaire generator tool permitted a highly user-friendly design, as confirmed by many favorable comments made by respondents at the end of the questionnaire. To avoid sampling issues (regarding variables like employment status and AIIC regions), the survey was addressed to the entire population, using the e-mail addresses published in the 2008 directory of members.

In line with the aim of a strict replication, we adopted Bühler's original criteria and response categories, with slight adjustments. In line with previous surveys among end-users, the focus was set on output-related features of quality rather than interpreters' personal qualifications. After careful consideration of any semantic implications, four items were used in a shortened or simplified form, including "logical cohesion (of utterance)" and "completeness (of interpretation)" as well as "(use of) correct terminology" and "correct grammar" (grammatical usage). In addition to Bühler's (1986) eight "linguistic" criteria and the "extra-linguistic" criterion of "pleasant voice," we included "lively intonation" and "synchronicity" to link up with the work of Collados Aís (1998, 2007) and Moser (1996), respectively. The four-point rating scale was modified to enhance its semantic consistency, replacing "irrelevant" by "unimportant," and the rating matrix also included a "no answer" option that was set as the default in the electronic questionnaire.

The survey design components that went beyond the study to be replicated included an open-ended question regarding the variability of quality criteria depending on meeting types as well as a range of items eliciting socio-

Figure 3. Quality criteria rated "(very) important" by 704 AIIC members (in percent)



professional background information on the respondents for subsequent correlational analyses. Moreover, the questionnaire comprised a second main component devoted to the issue of the interpreter's role, based on the assumption that quality-related priorities will be shaped by the way an interpreter perceives his or her role in the communicative interaction. This part of the survey could be discussed as yet another example of (partial) replication, with the study by Claudia Angelelli (2004) as its starting point. To keep the focus of this paper on quality, however, this dimension of our work, some of which has been reported elsewhere (Zwischenberger 2009), will not be discussed any further. Rather, we shall return to our replication of Bühler's study, the results of which, based on data from a total of 704 AIIC members (response rate: 28.5%) are shown in Figure 3.

As evident from Figure 3, the ratings given to the eleven output-related quality criteria by more than 700 members of AIIC with an average working experience of 24 years reflect a clear order of importance, as reflected in the percentages for "very important." The two criteria topping the list – "sense consistency with the original" and "logical cohesion" – are the same as in Bühler's study (and indeed in most other surveys using her list), though the percentages are somewhat lower and closer to those found by Kurz (1989) in

her study of user expectations (see Figure 2).

A similarly consistent pattern obtains at the lower end of the list, with Bühler's criteria of "pleasant voice" and "native accent" receiving the lowest percentages for "very important." In conjunction with the additional items of "lively intonation" and "synchronicity," one finds highly similar ratings for "pleasant voice" (which, at 28%, achieves the same percentage in the replication as in Bühler's original study) and for "lively intonation." This could indicate that respondents found it difficult to distinguish between the two. However, this had been anticipated at the design stage and motivated the decision to place the criterion of intonation ahead of "pleasant voice" in the list. The fact that the two criteria nevertheless received similar ratings can therefore be seen as confirming the conceptual proximity of "intonation," as vocal pitch movement, and "voice" as vocal sound quality in general, which also manifested itself very clearly in the work of the Granada group, who consequently embarked on a major effort at conceptual analysis in questionnaire-based "contextualization studies" (see Collados Aís et al. 2007).

Our replication also yields a clear-cut sequence of priorities for the criteria forming the poorly differentiated middle tier in Bühler's findings – terminology, fluency, grammar and completeness, all of which had been rated as "highly important" or "important" within a narrow range from 47 to 49 and from 49 to 51, respectively. As will be recalled, this lack of differentiation was what prompted Chiaro and Nocella (2004) to change from a rating to a ranking task with forced choice (i.e. no ties nor "no answer" option). Notwithstanding the potential merit of their methodological exercise, the present findings from the full-population survey among AIIC members show that a robust ranking can be established also with the original rating task.

Overall, the findings depicted in Figure 3 suggest a three-fold grouping of the eleven criteria used in our replication study. The first group is made up of the four top-ranking criteria, all of which were rated as "important" (or even "very important") by 99% of the respondents. These are "sense consistency with the original," "logical cohesion," "fluency of delivery" and "correct terminology." The second, middle-tier group is made up of five criteria (grammar, completeness, style, intonation, voice) that at least 86% of the sample regarded as "important" or "very important." Only two criteria – "synchronicity" and "native accent" – receive considerably lower ratings and thus make up the lower tier of the hierarchy. Whether these aspects of a simultaneous interpreting

performance are indeed and invariably of limited importance would have to be investigated among interpreters in studies modeled on the work of Collados Aís (1998). Substantial evidence that the pattern of priorities may shift depending on the type of meeting or assignment was collected with a subsequent item of our questionnaire, which suggests that future replications of Bühler's quality criteria should no longer ignore "hypertextual" variables.

Beyond the conference event or hypertext, broader socio-professional issues may also be at play in shaping interpreters' attitudes toward quality in simultaneous interpreting. Within the AIIC population, as represented very well by our large sample, hardly any socio-demographic variables were found to have an effect on the main findings regarding quality and role, suggesting a high degree of homogeneity, or shared views, presumably as a result of well-established traditions of training and professional socialization. However, the situation may be different if one were to look beyond the AIIC community and investigate the quality-related preferences of interpreters in regional/national markets. Clearly, this constitutes yet another case for replication – with extension, using the same instruments and techniques in a different population. This, too, has been undertaken as part of our project under the heading of national-level satellite surveys (For Germany, see Zwischenberger 2011). On the assumption, then, that the interpreting profession beyond the AIIC market may display some degree of sociocultural diversity, there is ample room for replication – in this case, partial, and perhaps even with cultural adaptation – as a way of broadening and deepening our knowledge about quality in simultaneous conference interpreting.

5. Conclusion

As this paper has sought to show, replication, defined as the systematic duplication of a previous study by other investigators, plays a vital role in the advancement of knowledge through empirical research. It is regarded as a cornerstone of scientific progress and, at the same time, a sobering reminder of what has been referred to as "the slipperiness of empiricism" (Lehrer 2010: 60). Findings may be skewed as a result of variability in the population (which applies to interpreters much more so than to laboratory mice, but even to the latter), which makes replication essential but replicability a fraught issue, not least for a highly context-bound and complex human performance such as

simultaneous interpreting. Such concerns are typically raised for experimental studies, and with good reason. As demonstrated here with reference to surveys on the topic of quality in conference interpreting, however, replication is no less needed or challenging in survey research. Even though repeating what someone else has done before seems more feasible than designing a study from scratch, and has therefore been suggested as a useful strategy for beginners (Gile 1990), replication, whether direct or partial, is by no means easy, and great care needs to be taken to get it right. While the implications of research findings in interpreting studies may be less consequential than those of clinical trials (in which replicability is often problematic), our knowledge about interpreting, and conclusions drawn for appropriate choices in professional practice and university-level training, should rest on a solid foundation. To the extent that we trust the standard methods of empirical science to deliver such evidence and findings, replication must be regarded, and regarded more highly, as a valuable approach to scientific research. Survey research on quality in conference interpreting offers us a number of interesting examples of replication studies, and a closer examination of them, with regard to errors and shortcomings as well as standard-setting achievements, enables us to learn valuable lessons and move forward by going back to what has been done before.

Acknowledgment

The author gratefully acknowledges the financial support of the Austrian Science Fund (FWF) for project P20164-G03 on "Quality in Simultaneous Interpreting" (QuaSI 2010).

References

AIIC (2005). AIIC: A statistical portrait (online). Retrieved from http://www.aiic.net/View Page.cfmpage_id=1906 on 11 February 2011.

Angelelli, C. V. (2004). *Revisiting the Interpreter's Role: A Study of Conference, Court, and Medical Interpreters in Canada, Mexico, and the United States.* Amsterdam/Philadelphia: John Benjamins.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua* 5(4): 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): 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. (2007). La incidencia des parámetro entonación. In Collados Aís, Á., E. M. Pradas Macías, E. Stévaux and O. García Becerra (eds.), *La evaluación de la calidad en interpretación simultánea: Parámetros de incidencia.* Granada: Comares, 159-174.

Collados Aís, Á., Pradas Macías, E. M., Stévaux, E. and García Becerra, O. (eds.) (2007). *La evaluación de la calidad en interpretación simultánea: Parámetros de incidencia.* Granada: Comares.

Denzin, N. K. and Lincoln, Y. S. (eds.) (2000). *Handbook of Qualitative Research* (2nd edn.). Thousand Oaks/London/New Delhi: Sage.

Dodds, J. M. and Katan, D. (1997) The interaction between research and training, in Gambier, Y., D. Gile and C. Taylor (eds.), *Conference Interpreting: Current Trends in Research.* Amsterdam/Philadelphia: John Benjamins, 89-107.

Gile, D. (1990). Research proposals for interpreters. In Gran, L. and Taylor, C. (eds.), *Aspects of Applied and Experimental Research on Conference Interpretation.* Udine: Campanotto, 226-236.

Gile, D. (1998). Observational studies and experimental studies in the investigation of conference interpreting. *Target* 10(1): 69-93.

Gile, D. (2005). Empirical research into the role of knowledge in interpreting: Methodological aspects. In Dam, H. V., J. Engberg and H. Gerzymisch-Arbogast (eds.), *Knowledge Systems and Translation.* Berlin/New York: Mouton de Gruyter, 149-171.

Hubbard, R. and Armstrong, J. S. (1994). Replications and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing* 11(3): 233-248.

Hubbard, R. and Vetter, D. E. (1997). Journal prestige and the publication frequency of replication research in the finance literature. *Quarterly Journal of Business and Economics*

36(1): 3-13.
Jolibois, S. (2010). *The Role of the Interpreter: An Overview.* MA dissertation, University of Burgundy.
Kopczyński, A. (1994). Quality in conference interpreting: Some pragmatic problems. In Snell-Hornby, M., F. Pöchhacker and K. Kaindl (eds.), *Translation Studies – an Interdiscipline.* Amsterdam/Philadelphia, John Benjamins, 189-198.
Kurz, I. (1989). Conference interpreting – user expectations. In Hammond, D. L. (ed.), *Coming of Age: Proceedings of the 30th Annual Conference of the American Translators Association.* Medford, NJ: Learned Information, 143-148.
Kurz, I. (1993/2002). Conference interpretation: Expectations of different user groups. In Pöchhacker, F. and M. Shlesinger (eds.), *The Interpreting Studies Reader.* London/ New York: Routledge, 313-324.
Kurz, I. (2001). Small projects in interpretation research. In Gile, D., H. V. Dam, F. Dubslaff, B. Martinsen and A. Schjoldager (eds.), *Getting Started in Interpreting Research.* Amsterdam/Philadelphia: John Benjamins, 101-120.
La Sorte, M. A. (1972). Replication as a verification technique in survey research: A paradigm. *Sociological Quarterly* 13(2): 218-227.
Lehrer, J. (2010, December 13). The truth wears off: Is there something wrong with the scientific method? *The New Yorker,* 52-60.
Mack, G. and Cattaruzza, L. (1995). User surveys in SI: A means of learning about quality and/ or raising some reasonable doubts. In Tommola, J. (ed.), *Topics in Interpreting Research.* Turku: University of Turku, Centre for Translation and Interpreting, 37-49.
Morrison, A., Matuszek, T. and Self, D. (2010). Preparing a replication or update study in the business disciplines. *European Journal of Scientific Research* 47(2): 278-287.
Moser, P. (1996). Expectations of users of conference interpretation. *Interpreting* 1(2): 145-178.
Pöchhacker, F. (2005). Quality research revisited. *The Interpreters' Newsletter* 13: 143-166.
Pöchhacker, F. (2011). Researching interpreting: Approaches to inquiry. In Nicodemus, B. and L. Swabie (eds.), *Interpreting Research in Theory and Practice.* Amsterdam/ Philadelphia: John Benjamins (in press).
QuaSI (2010). Quality in simultaneous interpreting (online). Retrieved from http://quasi. univie. ac.at on 11 February 2011.
Seleskovitch, D. (1996). Comment: Who should assess an interpreter's performance? *Multilingua* 5(4): 236.
Thompson, B. (1994). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. *Journal of Personality* 62(2): 157-176.
Vuorikoski, A.-R. (1993). Simultaneous interpretation – user experience and expectations. In Picken, C. (ed.), *Translation – the Vital Link. Proceedings. XIIIth World Congress of FIT* (Vol. 1). London, Institute of Translation and Interpreting, 317-327.
Zwischenberger, C. (2009). Conference interpreters and their self-representation: A world-wide web-based survey. *Translation and Interpreting Studies* 4(2): 239-253.

Zwischenberger, C. (2011). Quality criteria in simultaneous interpreting: An international vs. a national view. *The Interpreters' Newsletter* 15 (in press).

Zwischenberger, C. and Pöchhacker, F. (2010). Survey on quality and role: Conference interpreters' expectations and self-perceptions (online). Retrieved from http://www.aiic. net/ViewPage.cfm/article2510.htm on 11 February 2011.

Zwischenberger, C., Pöchhacker, F. and Kurz, I. (2008). Quality and role: The professionals' view (online). Retrieved from http://www.aiic.net/ViewPage.cfm/article2242.htm on 11 February 2011.

Author's e–mail address
franz.poechhacker@univie.ac.at

About the author
Franz Pöchhacker is Associate Professor of Interpreting Studies in the Center for Translation Studies at the University of Vienna. He was trained as a conference interpreter at the University of Vienna and the Monterey Institute of International Studies and has done freelance work as a conference and media interpreter since the late 1980s. He has conducted research on simultaneous conference interpreting as well as media interpreting and community-based interpreting in healthcare and asylum settings and published on general issues of interpreting studies as a discipline. His textbook on interpreting research, Introducing Interpreting Studies, has been translated into several languages, and he is co-editor of The Interpreting Studies Reader and of the journal Interpreting.