Toward a Document Evaluation Methodology: What Does Research Tell Us About the Validity and Reliability of Evaluation Methods?

Abstract—Although the usefulness of evaluating documents has become generally accepted among communication professionals, the supporting research that puts evaluation practices empirically to the test is only beginning to emerge. This article presents an overview of the available research on troubleshooting evaluation methods. Four lines of research are distinguished concerning the validity of evaluation methods, sample composition, sample size, and the implementation of evaluation results during revision.

Index Terms—Congruent validity, expert judgments, formative evaluation, Human–Computer Interaction (HCI), methodology, predictive validity, reliability, revision, usability testing, validity.

In the field of technical communication, the awareness has grown that usability evaluation research can be a fruitful way of monitoring, improving, and maybe even guaranteeing the user-friendliness and effectiveness of documents. Over the years, various empirical and analytical evaluation techniques have been added to the toolbox of technical communicators [1]–[4], and several recent handbooks describe in detail how to plan and conduct a usability test [5]–[10]. As a consequence, the development and validation of evaluation methods has become a potential area of scholarly research in technical communication. In 1989, Wenger and Spyridakis emphasized the importance of the traditional methodological criteria of validity and reliability for the practice of usability testing [11]. Last year, Hughes renewed this position and explored the ways in which these criteria can be adapted to qualitative evaluation research [12].

Two recent studies in the area of Human–Computer Interaction have painfully demonstrated that some of the current usability evaluation practices may be built on quicksand. The first was an extensive research project comparing the test results of eight teams of experienced usability professionals and one student team, who were all given the assignment to evaluate Microsoft’s Hotmail website [13]. The test results appeared to be highly inconsistent. The number of problems discovered by the teams varied between 10 and 150. Furthermore, 75% of all the problems found in the website were unique problems discovered by only one of the participating teams. Such findings draw attention to the validity and reliability of usability testing in practice. Should we not expect more agreement between the tests? What does a single
usability test say about the quality of the website? Given the variability among the usability teams, is it possible to distinguish effective from less effective test approaches? Although these results may clearly demonstrate the relevance of validity and reliability, the research in question does not give clues to answer these questions.

The second publication was a critical review by Gray and Salzman of five influential empirical studies comparing usability evaluation methods [14]–[16]. For each of these five studies, the research design and conclusions were carefully examined. According to these analyses, all the studies had so many methodological flaws (e.g., very small sample sizes, a lack of statistical testing, confounding variables, and contradicted conclusions) that it would not be justified to guide or justify usability practices on the basis of the results. Unfortunately, the latter is exactly what usability practitioners and handbook authors had been doing.

These findings sketch a worrisome overall picture for communication practitioners. Apparently, evaluation methodology matters, but we may know less than we thought we knew. In this article, we will present a more extensive overview of the methodological research available on the topic of document evaluation. We will not, as Gray and Salzman did, systematically focus on possible flaws in the design of existing research. Instead, we will discuss the relevant questions regarding document evaluation methods, the ways that these questions can be answered, and the main results. Our aim is twofold: (1) on the one hand, this article may serve as an overview of what is empirically known about document evaluation methods, and (2) it may be considered the start of a research agenda that can be filled in with new research initiatives.

The research discussed in this article is not limited to the field of technical communication. In our view, it would not make sense to draw strict lines between the evaluation research in various communication disciplines. In addition to methodological studies typically rooted in technical communication, we therefore also discuss research on the methodology of evaluating questionnaires, learning materials, user interfaces, and public information. The artifacts evaluated in these disciplines may show more or less resemblance to a document, but the general problem of investigating the pros and cons of evaluation methods will be quite similar. Of course, caution is needed when conclusions regarding, for instance, user interface evaluation are applied to the context of documents. But the same caution is needed when conclusions are translated from one type of document to another.

There is one important restriction in this article. Not all kinds of evaluation methods are included in our overview. In an earlier publication, we distinguished between verifying and troubleshooting evaluation functions [3]. Methods with a verifying function aim at obtaining overall impressions about document quality, which requires a quantitative research design. Methods with a troubleshooting function aim at detecting and diagnosing possible reader problems, which usually requires a qualitative and exploratory approach. The research discussed in this article focuses on evaluation methods with a troubleshooting function. Typical verifying approaches (e.g., readability formulas, the cloze test, and measuring task completion times) are set aside. In the Appendix, a brief description is given of the evaluation methods mentioned in this article.

Within this category of troubleshooting evaluation methods, the challenge of applying the traditional methodological standards of validity and reliability is often far from straightforward. As we will show, some intermediate steps may be required to put the validity and reliability of methods to the test empirically.

**Research Questions About Document Evaluation**

All research efforts into the validity and reliability of evaluation methods eventually go back to one practical question: how can we optimally benefit from possible evaluation activities during document design processes? In daily practice, usability professionals have to select an appropriate evaluation method, decide on the kinds of participants to recruit and on the sample size, and make sense of the large amount of user data they collect. All these aspects may affect the evaluation results and the usefulness of the efforts. Methodological research into document evaluation is meant to provide empirical support for the design of usability evaluation research. In this article, four clusters of research will be distinguished:

- **Method validity:** Which evaluation method can best be used to collect feedback on documents? Different evaluation methods may bring to light different types of feedback on a document. For a well-considered choice of an evaluation design, it is important to know which types of feedback methods do and do not yield, and what kinds of benefits can be reasonably expected.

- **Sample composition:** Which participants can best be recruited to collect feedback on documents? The background characteristics of participants—e.g., their educational level, age, or
personal involvement—may influence the types of feedback they give under a certain method. A careful evaluation design should not only address the methods to be used, but also the type(s) of participants to be selected.

- Sample reliability: What is the optimum sample size for an evaluation? The more participants are recruited for evaluating a document, the more feedback will be gathered, and the less the results will be biased by the peculiarities of individual participants. In practice, however, professionals must balance between time and cost considerations, and the exhaustiveness and stability of the results.

- Implementation and revision: How can the results of an evaluation be translated into effective revisions? The usefulness of troubleshooting evaluation must show in an improved effectiveness of documents. However, many user data are not self-explanatory and different experts may handle evaluation results differently during the subsequent steps to be taken from data collection to revision—i.e., the identification of problem areas, the diagnosis of problems, the judgment of the severity of problems, and the selection of suitable revision strategies [17].

The latter question is not a traditional methodological one, but it is equally important since we are dealing with the design and benefits of applied research methods. These four clusters of research form the basis of an extensive potential research agenda. The research agenda is not only extensive because there are so many methods and variations of methods to choose from, but also because the four questions may be strongly interrelated and because several context variables may be expected to be of influence. We will discuss the research available in each cluster below.

**Validity of Evaluation Methods**

In traditional methodology, validity refers to the degree to which a method is free of systematic bias. The main concern is whether a method really measures what it is intended to measure. Applied to the context of troubleshooting evaluation, the important issues are whether the problems found using a particular method correspond to problems readers have in real life and whether the method does not overlook problems that it is supposed to detect.

To some extent, all methods available have some “face validity.” For every data collection strategy, a more or less plausible rationale can be given. In practice, almost any attempt to systematically evaluate documents can be perceived as fruitful, regardless of the method used. The mere fact that a document receives intensive attention during an evaluation may already result in new and valuable insights. Hence, research is needed to find out the relative strengths and weaknesses of evaluation methods. The goal of validation research is to provide objective and empirical evidence to support or disprove assumptions based on method rationale and practitioners’ experiences.

The validity of evaluation methods can be investigated in several ways. Below, we distinguish between two main types of validation research:

- **Research into the Predictive Validity** of methods, investigating the relevance of the problems detected under a particular method.
- **Research into the Congruent Validity** of methods, focusing on the similarities and differences between the problems found under different methods.

**Predictive Validity** As the term predictive validity implies, the question addressed in this type of validation research is whether methods correctly predict the problems readers have with a document. Normally, the predictive validity of a research method is investigated by comparing its results to those of a “criterion instrument” that is known to yield valid results. In the case of troubleshooting document evaluation, such an instrument is usually unavailable. Instead, three types of research are carried out to shed light on the predictive validity of evaluation methods. These types of research and the main results will be discussed below.

**Revision-Based Studies:** The most current research design used to investigate the predictive validity of evaluation methods uses revision as an intermediate step. The research question is whether revision on the basis of evaluation results leads to significant improvements in the effectiveness of artifacts. To investigate this, the quality of original and revised artifact versions is experimentally compared, either in an independent-groups design with different groups of participants assigned to original and revised versions or in a motivated-choice experiment with one group of participants deciding on their preferences for either version.

Over the years, many revision-based studies have been carried out, together leading to the general conclusion that evaluation activities are worthwhile [18]–[34]. In a majority of the studies, the evaluation and revision appeared to have positive effects on the artifacts tested. For four of the studies, this overall conclusion cannot be specified to a particular method used. Jansen and Steehouder [28], Wright [30], and Schriver [32, pp. 444–462] based their revisions on an unspecified combination of expert analysis and reader feedback, while Allwood
and Kalén [31] used the input of three dissimilar types of evaluation data (i.e., underlining difficult passages, writing down questions, and think-aloud protocols) to revise a manual.

The remaining thirteen studies shed some light on the benefits of particular evaluation methods. For a more generic discussion of the findings, Table I gives an overview of the validation results for six types of methods:

- Text-focused methods (e.g., guidelines or checklists).
- Expert-judgment-focused methods (e.g., reviews by audience or genre experts).
- Reader-focused methods using task outcome (e.g., comprehension tests).
- Reader-focused methods using behavioral observation (e.g., think-aloud user protocols).
- Reader-focused methods using verbal self-reports (e.g., plus-minus method).
- Reader-focused methods using a combination of task outcome, observation, and self-reports (e.g., one-to-one evaluation).

For text-focused methods, only one study can be mentioned. Duffy and Kabiance investigated whether the application of traditional guidelines for readable writing would enhance the comprehensibility of texts [23]. Their conclusion was negative. Since the guidelines investigated were limited to basic word and sentence features, it is not justified to extend these findings to the usability heuristics and guidelines that have become popular since then.

For expert-judgment-focused methods, varying results were obtained in three studies. Weston et al. found no significant improvement for a revision of a textbook unit based on the comments of three subject-matter experts and three target-population experts [33]. Swaney et al. found that an evaluation by a document design team of four types of functional documents led to an improved effectiveness in half of the cases [29]. Finally, Davidove and Reiser discovered that a textbook revision based on the feedback of nine teachers helped students reach significantly higher test results [27]. In all of these studies, the usefulness of the expert feedback was compared with that of reader feedback. The results of Weston et al. and Swaney et al. suggest that reader feedback may be a more powerful tool to gear a document to the readers' needs; Davidove and Reiser—as well as Golas [25]—found no differences in effects between expert and reader feedback.

Two of the older studies into the benefits of reader-focused methods focused on the usefulness of task outcome data (e.g., comprehension test scores) to support revision [18], [20]. In both studies, significant improvements were found. Only one study addressed the benefits of observation and performance data. Swaney et al. found that an insurance policy was improved in a revision based on think-aloud protocols of readers who tried to use the information [29].

Four studies focused on the benefits of self-reporting reader feedback. The results of these studies were mixed. Ahlschwede established only minor improvements in a brochure based on an evaluation using a predecessor of the plus–minus method [19]. De Jong found that target readers preferred brochure versions that were revised on the basis of “plus–minus” reader feedback, and he established increased effectiveness in five of the six brochures he investigated [34, pp. 47–91]. Nathenson and Henderson [22] and Micklos and Bishop [24] established positive effects using other self-reporting evaluation techniques, in-text feedback questions, and the signaled stopping technique, respectively.

Finally, four studies investigated the benefits of a combination of the kinds of reader data described above, typically used in the context of instructional design, either as “one-to-one” or as “small-group” evaluation [21], [26], [27], [33]. All these studies established significant improvements by the empirically based revision.

One important reservation must be made, however, regarding the results described above. A potential weakness of revision-based validation research lies in the contribution of the reviser(s). A revision-based research design must ensure in one way or another that the effects found can be attributed to the evaluation results. In only seven of the revision-based studies, an attempt has been made to control the reviser’s role to some extent:

- Baker [20], Kandaswamy, Stolovitch, and Thiagarajan [21], and Golas [25] used a multiple-revision design, with different revisers using reader feedback to revise a document.
- Wager used a design in which the revisions were checked for appropriateness by a supervisor [26].
- Swaney et al. [29] and Weston et al. [33] used the “with/without” principle: By comparing the

---

**TABLE I**

RESULTS OF REVISION-BASED VALIDITY RESEARCH

<table>
<thead>
<tr>
<th>Type of Method</th>
<th>No Effect</th>
<th>Mixed Effect</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text-focused:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guidelines:</td>
<td>[23]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert-focused</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[33] [29] [27]</td>
<td>[27]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reader-focused:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>[18] [20]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reader-focused:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-reports</td>
<td>[19]</td>
<td>[34]</td>
<td>[22] [24]</td>
</tr>
<tr>
<td>Reader-focused:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination</td>
<td>[21] [26]</td>
<td>27 [33]</td>
<td></td>
</tr>
</tbody>
</table>
Problem-Assessment Studies: A second way to investigate the predictive validity of evaluation methods is by collecting additional data regarding the severity of the problems detected. The severity of reader problems can be determined by assessing their likelihood (To what extent is it plausible that target readers will have the problem concerned?) and their impact (To what extent is it plausible that the problem concerned will be a threat to the effectiveness of the document?) [35], [36].

The severity of problems detected can be determined in two ways. The first possibility is to design a test to check the occurrence of all problems detected in a new and preferably larger sample of target readers. This type of research would focus on the likelihood of problems and does not necessarily address their impact. We have not come across research using this approach.

A second possibility is to have experts (or experienced readers from the target audience) judge the severity of the problems detected. A few examples of this approach can be found, but most of these are not primarily aimed at validating a particular method of testing, but at differentiating between important and unimportant feedback for other purposes [35]–[38].

De Jong used assessment by experts as an additional source of information about the validity of the plus-minus method for pretesting brochures [34, pp. 93–105]. The results appeared to be hard to interpret, due to a lack of agreement among the participating experts. The experts generally agreed that a substantial part of the plus-minus results referred to important reader problems (on average 37% of the reader problems), but they did not agree at all about which reader problems were the most important. As a result, the majority of the reader problems received a meaningless neutral mean score, somewhere in the middle of the five-point scale used. This may be an important drawback of this type of validation research.

Error-Detection Studies: A third way of investigating the predictive validity of evaluation methods is to make use of the “error-detection” paradigm. This approach is often used in theoretical comprehension-monitoring research, aimed at establishing whether readers are aware of the comprehension problems they encounter [39], [40]. In this research, readers are presented with a manipulated text deliberately containing certain types of problems. The question of interest is whether the readers will indeed notice the problems included in the text.

Four error-detection studies can be found focusing on the validity of evaluation methods. Pander Maat investigated the effectiveness of two variations of the plus-minus method in detecting several types of problems in a drug information leaflet: (a) problems with medical or pharmacological terms, (b) problems with quantifiers, (c) problems with references to medical conditions, (d) problems of missing information, and (e) internal inconsistencies in the text [41]. The prediction rates did not exceed 50% for any of the five categories; very low prediction rates were found for problems with quantifiers (6%) and missing information (2%).

In the context of questionnaire pretesting, a cluster of studies used the error-detection approach in a questionnaire containing five types of problems: (a) loaded questions, (b) double questions, (c) ambiguous questions, (d) inappropriate vocabulary, and (e) missing response alternatives. Using three evaluation approaches (telephone interviews, face-to-face interviews, and think-aloud protocols), Hunt, Sparkman, and Wilcox found very low prediction rates for the first four categories (between 3% and 15%), and somewhat higher rates (33%) for the missing response alternatives [42].

With relatively highly educated participants, Diamantopolous, Reynolds, and Schlegelmilch [43] established considerably higher detection rates (between 16% for double questions and 48% for inappropriate vocabulary). Finally, Reynolds and Diamantopolous compared the effectiveness of a personal (face-to-face) and an impersonal (written) evaluation method to detect the five types of problems, again with relatively highly educated participants [44]. They established that the personal evaluation method was significantly more effective than the impersonal method, with average detection rates of 32% and 20%, respectively. The difference can be attributed mainly to the problems of inappropriate vocabulary.

The error-detection paradigm has the advantage that the research design is more focused and simple than with the other two types of research. However, the approach is not entirely compatible with the basic assumptions of troubleshooting evaluation (i.e., that evaluation may result in new and surprising insights into reader problems). Besides, in principle, additional evidence is needed to show that the problems included are in fact real reader problems. Pander Maat carried out an extra check on the problems he included in the leaflet, which caused him to delete three of the fourteen problems [41]. In the other studies mentioned above, the authors state that the problems were
“quite blatant,” which may be true from a questionnaire designers’ point of view, but they seem to be more problematic for the researchers and analysts than for the future respondents [42]–[44]. A peculiarity of all the error-detection studies so far has been that they do not report on the percentage of all problems included detected by at least one of the participants. In this respect, the error-detection studies are, despite their claims, not really geared to the validation of troubleshooting evaluation methods.

**Conclusion**: The available research into the predictive validity of evaluation methods confirms the basic idea that a troubleshooting evaluation is worthwhile. However, little is known yet about the predictive validity of particular evaluation methods. The three current types of validation research all have their pros and cons, and they may complement each other very well. Revision-based studies can be used to obtain an overall impression of the possible benefits of evaluating documents using a certain method. Problem-assessment studies may, in principle, be useful to differentiate between important and unimportant problems. Error-detection studies may be a useful complement because it departs from a set of problems that an evaluation method, according to the researcher, should have brought to light instead of what the method has in fact revealed.

**Congruent Validity** Research into the congruent validity of evaluation methods focuses on similarities and differences in the evaluation results. Three aspects may be involved in the comparison of methods:

- The numbers of problems detected.
- The attention paid to various types of problems “evaluation standards” [39].
- The degree of overlap in the problems detected.

The number of problems detected is a rough overall indicator of a method’s yield. The underlying assumption of “the more the better” may be more justified for observational methods focusing on usability problems during task performance than for self-reporting evaluation methods with a broader focus.

Evaluation standards may reveal tendencies in the problem types detected using a certain method. The research focuses on the distribution of problems over various predetermined problem categories, which can be—and have been—defined in various ways, depending on the type of documents evaluated and the methods used. In the context of questionnaire pretesting, Presser and Blair distinguished between “respondent semantic” (problems participants have interpreting a question), “respondent task” (problems participants have answering a question), “interviewer” (problems interviewers have asking a question or recording the response), and “analysis” (problems the researcher has interpreting the results) [45]. In another study, Oksenberg, Cannell, and Kalton chose a similar, though somewhat more elaborate, approach [46]. In several other studies, variations on the following typology of problems were used [34, pp. 47–91], [47]–[51]:

- **Comprehension**: Readers experience problems due to lack of clarity and problems with the applicability of information, or with difficult syntax or choice of vocabulary.
- **Acceptance**: Readers disagree with factual information, value judgments, or advice given in the document.
- **Appreciation**: Readers simply prefer another formulation but do not mention a problem with comprehension or acceptance.
- **Structure**: Readers have problems with the ordering of information in textual units or with the signaling of the structure (e.g., headings or cross-references).
- **Relevance**: Readers claim that certain information should not be included in the document or could at least be covered in less detail.
- **Completeness**: Readers ask for more information about the topic or for more elaboration on a certain point.
- **Graphic design**: Readers are critical of the document’s layout or illustrations.
- **Correctness**: Readers notice a violation of syntax, spelling, punctuation rules, or text conventions.

The degree of overlap represents the most detailed level of similarities and differences between methods. This type of research may lead to conclusions about the extent to which method A predicts the results of method B, or the extent to which methods A and B complement each other. However, the matching of reader problems detected by means of different methods may be a serious problem in practice, especially if both methods yield entirely different types of feedback (e.g., in the case of a comparison of usability test data and the results of a heuristic evaluation). Lavery, Cockton, and Atkinson notice a lack of rigor in this respect in current research efforts, and therefore propose the use of structured usability problem reports to facilitate the comparison of individual problems [52].

Ideally, research into the congruent validity of methods pays attention to all three aspects of comparison. In practice, however, this is seldom the case. Due to the variety of methods, it is impossible to present a complete and detailed overview of the results. Instead, we will discuss the following clusters of research:

- Research comparing expert and reader feedback.
• Research comparing observational data and self reports.
• Research comparing individual and group data collection.

Expert versus Reader Feedback: The question as to whether experts—with or without the help of heuristics or other facilities—provide different feedback on documents than readers has been extensively investigated [32, pp. 444–462], [35], [41], [45], [48], [49], [53]–[59]. The results of one of these studies are hard to interpret. Weston concluded that experts and readers provide different feedback on learning materials, but the experts and readers involved were asked entirely different questions about the materials [55]. As a result, the observed differences cannot be unambiguously attributed to the participants in the evaluation. Other studies had a more univocal approach, either asking experts to predict the results of a reader-focused evaluation or giving experts and readers comparable evaluation purposes.

Several researchers have investigated to what extent experts can predict the results of reader-focused evaluation. Rothkopf found a strong negative correlation between the observed effectiveness of an educational program and the predictions by experts [53]. In the context of questionnaire pretesting, Presser and Blair found a moderate degree of overlap between the results of expert evaluation and several reader-focused techniques (Yule’s $Q = 0.36$) [45]. Three other studies established that individual experts—professional writers, intermediaries, or subject-matter experts—were not able to predict more than 15% of the reader problems detected in a “plus–minus” evaluation, regardless of the severity of the problems [41], [48], [49]. Mack and Montanitz came to the same conclusions—with rates of 18 to 22%—for software designers and usability professionals who tried to predict the usability problems in an interface [58]. Schriver found higher prediction rates for a document design team in an iterative usability evaluation of manuals: 50% of the user problems detected in the original manual were predicted, but only 19% in the revised manual [32, p. 454]. Dieli found even higher prediction rates for ten document designers evaluating a manual (72% of all reader problems), but she compared problem areas instead of problem descriptions, which may have distorted her results [54]. Across the board, the conclusion seems justified that the feedback of real readers often contains many surprising new insights for expert evaluators.

Another side of the prediction question concerns the “extra” feedback that experts produce, feedback that would not be obtained using a reader-focused evaluation technique. Even when experts are given the explicit task to restrict their evaluation to the problems readers would have, between 40% and 70% of their predictions does not correspond to problems mentioned or experienced by target readers [41], [48], [49], [54], [59]. Lentz and de Jong found a significant difference in this respect between types of experts: professional writers were considerably more likely to put forward new problems than intermediaries [49]. So far, there have been no studies investigating the validity of the “false alarms” produced by experts.

Several studies have investigated the effects of recently developed approaches, guidelines, or checklists to support expert evaluation [35], [54], [56]–[59]. Compared to unguided expert evaluations, Dieli reported higher prediction rates (85%) when experts used the “revision filters” she developed, using the aggregated results of reader feedback, and remarkably lower prediction rates (35%) when experts used traditional writing guidelines [54]. Apparently, the kind of guidance given to experts may seriously affect their way of evaluating a document, and the basic idea of usability heuristics can be a fruitful one. The other research available ended in varying results. Two studies—both with very small numbers of participants in the reader or user evaluation—were encouraging. Lentz and Pander Maat found that experts working with their “functional evaluation checklist” predicted all the problems in a brochure that were detected in a small-scale evaluation using the plus–minus method [56]. Nielsen found that a heuristic evaluation by experts revealed more than 80% of the interface problems detected in a usability test [35].

On the other hand, Desurvire established that experts, some working with heuristics and others with a cognitive walkthrough, never predicted more than 50% of the interface problems found in a usability test and missed the serious problems more often than minor annoyances [57]. Mack and Montanitz discovered prediction rates of 13% to 15% for experts using a cognitive walkthrough [58].

John and Marks approached the same question from the other side. They investigated whether the problems put forward by experts using one of five supporting tools—i.e., claims analysis, cognitive walkthrough, GOMS, heuristic evaluation, and user action notation—were confirmed by a usability test [59]. More than half of the predicted problems did not correspond to a test result. In Nielsen’s heuristic evaluation study, the percentage was even higher (58%) [35]. The authors of both studies seem to disagree on the interpretation of this finding. John and Marks categorize the unconfirmed problems as “wasted effort,” while Nielsen argues that these extra problems may be very useful.
Systematic differences between experts and readers can be discovered by analyzing the “evaluation standards” used. De Jong and Lentz found that, unlike readers, experts paid more attention to presentation issues and less to the information offered [48], [49]. Dieli also established a relatively strong attention among experts to stylistic issues [54]. Pander Maat observed that readers focused relatively strongly on comprehension problems in a patient information leaflet, and experts provided more feedback on the acceptability, structure, and completeness of the information provided [41]. He also describes how experts and readers may put forward different problems within the problem categories. For instance, with regard to acceptability, the readers mostly expressed their mistrust in the information given or commented on the feasibility of advice, whereas the experts provided more factual rectifications. The evaluation standards experts use are closely related to the role they see for themselves. In an ethnographic study of document reviews, Raven observed that expert reviewers had difficulty in sticking to their evaluation task; the subject-matter experts could not refrain from making a lot of stylistic comments [60].

All in all, the conclusion must be drawn that expert-focused and reader-focused evaluation are two different things. Approaches such as heuristics and cognitive walkthroughs may help experts to evaluate a document from the users’ point of view, but the empirical evidence to support this claim is scarce. A promising approach lies in the education of writers. Studies by Schriver and Couzijn proved that writers—college students and high-school students, respectively—can learn from being systematically confronted with reader feedback [61]–[63]. In our view, however, it is not sensible to consider expert evaluation only as a surrogate for reader feedback. Experts can make their own contribution to the quality of documents and supplement the results of reader-focused evaluation because they have a better overview of the goals of the document and of alternative design options or because it will be impossible to evaluate every aspect of a document from all points of view in a user test.

**Observational Data Versus Self-Reports:** Another topic regarding the congruent validity of evaluation methods deals with the distinction between observational data (such as think-aloud user protocols) and self-reporting approaches (such as the plus–minus method). In principle, there is a tradeoff between the two types of methods. On the one hand, observational data are supposed to be superior when it comes to identifying usability problems. Self-reporting participants may not realize that they have a comprehension problem and may decide not to report a problem because of the impression it would make (and actively search for other problems to please the researcher), or they may forget about the problems they experienced before reporting them to the researcher. On the other hand, self-reporting participants may be expected to give feedback on a broader range of characteristics. Some of the studies we encountered provide empirical support for these assumptions.

Although not overwhelming, five studies provide some empirical support in favor of the observational data. Henderson et al. compared (retrospective) think-aloud protocols to logged data, questionnaire data, and interview data [64]. The think-aloud method revealed more problems than the logged data and the questionnaire, but there were no significant differences with the interviews. Allwood and Kalén compared the yield of think-aloud protocols to the number of problems detected with participants underlining difficult passages in a manual or writing down questions that arose during reading [31]. Only marginally significant differences were found between the “think-aloud” and “underlining” condition. In a small-scale study, Medley-Mark and Weston compared the problems identified by an individual participant thinking aloud, a pair of participants discussing the problems they encountered, and a small group of three participants who work silently without intervention and were interviewed afterwards [65]. The think-aloud participant proved to be most productive, the small group of three identified the fewest problems, as well as the fewest problems at a detailed level. In the context of questionnaire pretesting. Bischoping established that a debriefing of interviewers in a focus group discussion did not correspond very well to the problems found in the behavioral protocols made of interviewer and respondent behavior [66]. Finally, Gillham and Buckner found a discrepancy between the behavior of people using a multimedia encyclopedia on CD-ROM and their ranking of important features [67]. Although participants spent a lot of time playing with the CD-ROM, they did not rank the importance of multimedia particularly high. Similarly, although participants spent little time reading textual information, they stated that the text content was a crucially important factor.

Another study, by Sienot, confirmed that self-reporting evaluation methods reveal a broader range of problems than methods based on behavioral observation [50]. Sienot compared the plus–minus method and think-aloud user protocols for the evaluation of websites. The plus–minus method revealed significantly more distinctive problems than the think-aloud protocols, which must be
particularly attributed to the category of appreciation problems.

In the same experiment, Sienot also investigated whether the evaluation results would be influenced by the assignment given to the participants. Half of the participants visited the website with specific tasks, the other half surfed the website freely. The participants without tasks brought to light more problems, particularly concerning the relevance and completeness of the information. On the other hand, they reported significantly less comprehension problems than the participants who had been given tasks. This last finding more or less corresponds to the results of a comparison of “user protocols” (think-aloud protocols with tasks) and “reader protocols” (think-aloud protocols without tasks) by Dieli [54]. According to Dieli, the “user protocol” participants focused mainly on access and use, whereas the “reader protocol” participants restricted their evaluation to meaning.

**Individual Versus Group Data Collection:** De Jong and Schellens conducted a study that compared the way the plus–minus method is used by individual participants with the way it is used by focus groups [47]. On the basis of the general literature on focus groups, several assumptions can be made about the dynamics of the group discussion (e.g., participants influencing each other’s opinion, holding back comments, or building further on the comments of others) many of which were indeed observed in this study. The aforementioned study by Bischoping also revealed that the participants did not put forward all the questionnaire problems they had written down during the group debriefing session [66].

The main question in de Jong and Schellens’ research, however, concerned the way these dynamics might influence the problems detected. There were no differences regarding the number of problems detected per participant. However, there were differences regarding the types of problems detected. The focus-group approach urged participants to provide more feedback on the brochure as a whole and less on the word level. In addition, participants in the focus group condition focused more on acceptance problems in the brochure, and less on comprehension problems. On the basis of these results, the focus-group approach seems particularly useful when the persuasiveness of a document is deemed to be critical.

**Other Studies:** In addition to the clusters of research described above, other comparisons have also been made. In the context of questionnaire pretesting, Oksenberg, Cannell, and Kalton investigated whether special probing questions would result in more problems being detected, compared to the coding of interviewer and respondent behavior [46]. The answer was affirmative for one type of probing question only, that which further explored participants’ comprehension.

Smilowitz, Darnell, and Benson compared the results of laboratory usability testing with a formal beta test and a “forum test” with participants using the software and discussing it on the bulletin board [68]. The forum test proved to be the least effective approach. The lab test and the beta test revealed a similar number of user problems, but the lab test brought to light relatively more severe problems. As there was little overlap between the problems detected in the laboratory and the problems detected in the beta test, the two methods can complement each other.

In the future, more research into the congruent validity of methods is needed to provide a better insight into the effects of methods on feedback gathered.

In addition to the comparisons between methods, as described above, within-method comparisons focusing on the effects of method variations will become very important to help fine-tune evaluation designs.

**Conclusion:** The available research into the congruent validity of evaluation methods demonstrates that the method of evaluation makes a difference for the feedback collected about a document. This seems particularly the case for the comparison of expert feedback and reader-focused evaluation methods. Also, within the category of reader-focused evaluation methods, differences can be found, for instance, between behavioral observations and participants’ self-reports, or between individual and focus-group approaches.

**SAMPLE COMPOSITION**

The validity question of an evaluation design not only depends on the type of data collection method used, but also on the characteristics of the participants involved. Another line of research therefore focuses on the ideal participants to recruit for a document evaluation. Only three studies can be mentioned in this context.

In a very small-scale validation study, Wager investigated the influence of participant aptitude on the feedback gathered in one-to-one evaluations of educational materials [26]. She compared the results of three low-aptitude, three high-aptitude, and three mixed-aptitude students. She found that the low-aptitude students identified more basic problems in the module (e.g., with the vocabulary used), while the high-aptitude students were able to pinpoint inadequacies and to provide supplementary instructions. The mixed group gave the greatest variety of feedback. In a follow-up experiment, Wager established that a revision on the basis of the mixed-group
results was more effective than the revisions based on high-aptitude or low-aptitude results.

Diamantopoulos, Reynolds, and Schlegelmilch carried out an error-detection experiment to explore whether participants’ expertise in the field of questionnaire design and their prior knowledge on the questionnaire’s topic have an effect on the number of problems they detect [43]. The expertise variable was manipulated by selecting two groups of participants in the study; one group that had followed a course on questionnaire design, and one group that had not. The knowledge variable was manipulated in a simple and elegant way. One group received a questionnaire about British political affairs, and one group received exactly the same questionnaire about the Spanish situation. They found a significant effect for both independent variables without interaction effects. Questionnaire expertise was helpful to detect problems with ambiguous questions, leading questions, and missing response alternatives; prior knowledge appeared to facilitate the detection of double questions and missing response alternatives.

However, in a follow-up study by Reynolds and Diamantopoulos, using face-to-face and written evaluation approaches, the positive effect of prior knowledge was not confirmed [44]. Prior knowledge only proved to be helpful for detecting problems with double questions. There were no interaction effects between the participants’ prior knowledge and the method of evaluation.

**Conclusion** The few studies available suggest that the types of participants recruited may indeed affect the evaluation results. However, given the small scale of Wager’s study and the specific context of the other two studies, very little can be said about the exact ways in which participants’ background characteristics may influence the results. In future research, it seems important to differentiate between behavioral-observation and self-reporting methods. In the case of behavioral-observation methods, a strong relationship may be assumed between the characteristics of target readers and of participants recruited in a test. In the case of self-reporting methods, several intermediate factors may complicate the relationship between target readers and participants:

- **Metacognitive skills:** the degree to which readers are able to reflect on their own process of using and comprehending the text [39], [40].
- **Evaluation standards:** the characteristics participants are able to involve in their evaluation of a document [39].
- **Metalinguistic skills:** the degree to which readers are able to reflect on and discuss the quality of documents.

These three factors may be closely related to the educational level of participants. On one hand, feedback of participants with a lower level of education can be useful because it bridges the widest possible gap between writer and target readers. On the other hand, highly educated participants may provide more exhaustive and rich feedback on documents, from which the readers with a lower level of education may also benefit. Apart from educational level, prior knowledge and involvement of participants seem to be other factors that are worthwhile to investigate.

**SAMPLE RELIABILITY**

So far, the research discussed has dealt with the validity of the results: both method and sample composition may affect the usefulness of the feedback collected. Another line of research focuses on the traditional concept of reliability, i.e., the extent to which the results of an evaluation are stable. This is closely related to the sample size used. In the case of troubleshooting evaluation methods, the relationship between validity and reliability is not as strict as in quantitative research. Even if an evaluation yields highly unstable results, it may still provide valuable clues for improving a document. No serious evaluator is likely to be satisfied with a random collection of feedback on a document. Other than in quantitative research, however, there are no clear-cut statistical procedures to check the reliability of the evaluation data collected.

An approach used to investigate the reliability of an evaluation design is to conduct two evaluations and compare the research. Bischoing did this for the interviewer debriefing sessions in questionnaire pretesting [66]. She compared the problems put forward in two sessions (one group with six interviewers and one with nine). The results were mixed, depending on the type of problem involved. The two groups reported entirely different problems regarding the reading of questions and reached very high agreement on the topics of interruptions by respondents and inappropriate questions. On other problem types, there was moderate agreement. Presser and Blair used the same approach for various methods of questionnaire pretesting [45]. They found a rather high degree of overlap in problem identifications under the behavior coding method and much less reliable results under the other three methods (i.e., interviewer debriefing, think-aloud protocols, and expert evaluation).

A less static and probably more informative research approach exploring the relationship between sample size and yield uses a so-called “Monte Carlo” analysis [34, pp. 107–113], [35], [37], [38], [48], [49], [69]. From the total sample of participants, many
random subsamples of increasing sizes are taken, and for each sample size, the mean number of distinct problems detected is computed. The resulting graph can be extrapolated using a mathematical formula. In general, the results of a Monte Carlo analysis reflect the principle of diminishing gains per participant as more participants are included in the evaluation. The question of interest is how large the sample needs to be before no new problems (or only very few) are added to the list. The main results of the “Monte Carlo” studies available are listed below.

For behavioral observation methods, both Virzi and Nielsen came to the conclusion that five participants will reveal 80% of the usability problems in an interface [37], [69]. Lewis, on the other hand, found that five participants only revealed 55% of the usability problems [38]. While Virzi’s research is often cited to justify testing with very small samples, the issue actually remains unsolved.

With regard to self-reporting method with much broader feedback possibilities, de Jong concluded from an analysis of “plus–minus” results on four brochures that even sample sizes of 30 to 40 participants would not provide an exhaustive list of problems [34, pp. 107–113]. Though the principle of diminishing gain was visible in the resulting graphs, the possibility of adding new problems appeared to be almost endless. To further investigate the stability of the results, de Jong conducted a variation of the Monte Carlo procedure, drawing two subsamples from the total sample of participants and then computing the average amount of overlap in the problems detected. For an overlap of 60%, a sample of 22 to 36 participants would be needed. For an overlap of 80%, the required sample size would even be 40 to 65 participants.

For expert evaluation, the results vary. Nielsen discovered that five experts using heuristics to evaluate an interface would reveal 75% of the problems [35]. On the basis of cost–benefit considerations, he recommends the use of three to five evaluators. De Jong and Lentz, on the other hand, found that the graph of unguided experts evaluating brochures was hardly curved [48], [49]. The great majority of the problems mentioned by experts appeared to be unique detections, not shared by any of the colleagues.

**Conclusion** All in all, very little is known yet about the required sample size in document evaluation. It seems plausible that a sample size must be larger for self-reporting evaluation methods than for behavioral observation methods. Even in the latter category of methods, caution is needed when it comes to following Virzi and Nielsen’s attractive recommendations about sample size. Special caution is also necessary in regard to expert evaluation. From de Jong and Lentz’s studies, the conclusion must be drawn that experts simply do not agree regarding the problems they detect in a document. A single expert review cannot be considered a reliable evaluation approach; a review by ten or more experts, on the other hand, can easily result in an unmanageable set of problems.

**Differences Between Revisers** Two studies explored to what extent the revision of a document on the basis of evaluation results may depend on differences between individual revisers. Kandaswamy, Stolovitch, and Thiagarajan asked four teachers to revise educational materials twice by using one-to-one and small-group learner feedback [21]. In an experimental study, they compared the effectiveness of their achievements and established significant differences. All revisions appeared to be improvements, but the effectiveness varied among the revisers. However, the revisers in this study did not use the same reader feedback, which makes it hard to distinguish between their specific contributions and that of the reader feedback they had at their disposal.

**Implementation and Revision**

The lion’s share of the research available on document evaluation focuses on data collection. So far, very few studies have paid attention to the equally important problem of interpreting evaluation results and translating them into effective revisions. They first and foremost demonstrate how little is known yet about the process of revising on the basis of external feedback, and how potential deficiencies in this process may undo many of the efforts to collect valid and reliable evaluation data.

In our view, this is a fascinating and challenging new area waiting for serious research initiatives.

Two lines of research involve the entire process of revising on the basis of evaluation results, i.e., the differences between revisers using evaluation results and the role of evaluation results during revision. The other lines of research focus on specific aspects of using evaluation data identifying reader problems, judging the severity of problems, and selecting a revision strategy [17].
findings in more general research into revision [71].

Use of Evaluation Results Three studies addressed the role evaluation results play during the revision of an instructional text unit. Weston et al. asked four instructional designers to revise a unit using varying kinds of input [33]. Reviser 1 received reader feedback (learner comments and test results) and expert feedback (from subject-matter and target-audience experts). Revisers 2 and 3 received reader feedback and expert feedback only, and reviser 4 received no revision input. Despite the availability of evaluation results, revisers 1, 2, and 3 based 62% to 78% of their revisions on their own experiences and judgments. When using evaluation results, the experts in this study seemed to be more inclined to use reader feedback rather than expert feedback.

Le Maistre and Weston came to similar conclusions in a study with eight instructional designers working on two different texts [72]. The participants were provided with the same types of evaluation results as described above. On average, 80% of the revisions were based on the revisers' own insights, rather than on the evaluation data. In the remaining 20% of the revisions, a significant preference for evaluation data was found: (a) learner comments (12%) and target population expert data (9%), (b) subject-matter expert data (4%), and (c) test scores (2%). The total percentage of input used exceeds 100%, due to overlap in feedback between different sources.

Raven carried out an interesting ethnographic study that followed the revision process of two technical writers who were confronted with review comments by subject-matter experts [60]. Asked about their motives, both writers stated that the great majority of their revisions had not been made in order to improve the documents but to "maintain a good interpersonal relationship." Apparently, the writers felt more or less obliged to use the reviewers' comments.

Some tentative conclusions may be drawn from these studies. First, reader feedback seems to be more persuasive to revisers than the comments given by other experts. Second, the impact of evaluation results during the phase of revising documents may be limited in practice. Therefore, it seems questionable whether revisers optimally benefit from the data they work with.

Identifying Reader Problems In research reports and in handbooks on evaluation, authors seem to suggest that the identification of reader problems follows automatically from the data collected. In practice, this is not true. For instance, with think-aloud user protocols, the researcher or reviser has to identify the problem areas in the participants' process of using the document. Issues of validity and reliability are relevant here: what are effective and justified signals in user protocols for problem areas, and to what extent can different revisers be expected to identify the same problem areas in a user protocol?

So far, only one study investigated the identification of reader problems. In the context of questionnaire pretesting, Bolton developed and tested a way of automatically coding think-aloud protocols for the occurrence of reader problems [73], [74]. She distinguished various verbal and nonverbal cues in the protocols and tried to connect them with different problem types. The verbal cues consisted of lists of current expressions indicating that the respondents had difficulty answering a question; the nonverbal cues consisted of questions, pauses, broken utterances, and unintelligible utterances. A factor analysis confirmed the relation between eight of the nine cues and problem types. Bolton also compared the results to the problems identified by human coders. The overall consistency between the coding methods appeared to be low. Human coders have the advantage that they may include the adequacy of answers in their observations. On the other hand, automatic coding may lead to the identification of problems that human coders fail to detect.

Judging the Severity of Problems Another research issue concerns the judgment of the severity of problems. Revisers working under time constraints have to prioritize the most important problems in a document. Even when there is enough time, it is important that experts can reliably distinguish the "real" reader problems from the nonproblems. This goes for all pretest methods, but particularly for those using self-reporting data. To please the researcher, participants may be inclined to mention problems they do not really experience as such. The main research topic here is the inter-rater reliability among different experts in their severity ratings of problems.

Several studies indicate that individual experts cannot reliably judge the severity of problems. In the context of interface design, Nielsen asked experts to rate the problems found in a heuristic evaluation [35]. The mean correlation between the severity ratings of individual experts was only 0.24. Lentz and de Jong found a mean correlation between expert ratings of 0.17 [49]. In a study with four brochures, de Jong found correlations between 0.12 and 0.29 [34, p. 103].

Virzi suggests that there may be a strong relationship between problem severity and frequency [37]. In his study, the more severe user problems were easier to detect than the less severe ones, but this result may have been biased
because the severity coders knew the frequency of the problems in the test. In contrast, both Lewis and de Jong found that there was no relationship at all between severity and frequency of problems, with mean correlations of 0.06 and 0.08 ([38], [34, p. 101]).

Hassenzahl investigated the possibility of having interest-group members judge the severity of usability problems [75]. He compared their severity ratings with empirical data about the impact of the problems and found no correspondence. In a follow-up questionnaire, he explored to what extent several factors contributed to the severity ratings. The error-handling time and the psychological costs that problems would cause appeared to play an important role in the severity ratings. A dysfunctional factor was the participants’ inability to imagine the consequences of a problem. When the participants had difficulty imagining a problem’s consequences, they gave it a higher severity rating.

Lentz and de Jong explored whether the agreement among experts, i.e., professional writers, can be facilitated by (controlled) consultation [36]. In a Delphi study, experts rated the likelihood and severity of problems in three subsequent rounds. In the first round, they received no input. In the second round, they received an overview of the arguments given by the others to support their ratings. Finally, in the third round, they were given a short list with the main arguments and an indication of the rough tendencies in the mean ratings. In the first round, the experts strongly disagreed. Of the two subsequent rounds, only the third resulted in an improved degree of agreement among experts, though they still disagreed considerably. An analysis of the arguments provided by the experts demonstrated that their criteria for judging reader problems may have differed in many respects. The experts appeared to have trouble restricting their ratings to the likelihood and impact of problems, and often judged problems by exploring the benefits of alternative formulations. Some of the arguments reflected the difficulties some of the experts had empathizing with the readers, even when confronted with their problems.

Selecting Revision Strategies Concerning the selection of suitable revision strategies to fix reader problems, only one study can be mentioned. Schellens and de Jong asked five professional writers to revise brochure fragments containing 18 reader problems concerning the acceptability of factual information, the acceptability of normative statements, and stylistic appreciation [76]. In a follow-up questionnaire, the experts were asked to judge the quality of all revisions. The results confirmed some of the tendencies described above: the experts chose very different revisions, did not agree about the quality of the revisions, and often preferred their own solution. Interestingly, however, relationships were found both in the revision strategies chosen and in the ratings of revisions between types of problems and types of revision. In the case of problems of factual acceptance, the addition of extra information was a favorite revision strategy. In the case of problems of normative acceptance, the revisers often substituted the problematic statement. In the case of appreciation problems, the passage in question was often deleted or substituted. The linking of problem types and revision strategies may offer fruitful perspectives for future research.

Conclusion The results of a troubleshooting evaluation will only be worthwhile if they can be used effectively during the phase of revision. Very little is known, however, about the way revisers work with evaluation results and about effective and less effective strategies during revision.

The activities of identifying and judging problems put forward by experts or readers may require the development of new professional skills. The scarce research available mainly shows that there is work to be done in this area and provides some interesting starting points for future research.

Discussion

The substantial number of studies referred to in this article illustrates that evaluation methodology has increasingly received research attention among scholars and practitioners. However, it also veils how little is known about the enormous variety of methods available. More research is needed, and, above all, more systematic research programs on the topic of document evaluation are needed. The typology of research used in this article may be helpful to find the relevant questions to be investigated and to structure future research efforts.

With regard to method validity, it is important to distinguish predictive and congruent validity and collect data on both aspects. It may be very informative to explore the differences and similarities between the yield of different evaluation approaches (congruent validity), but there are also questions to be answered about the usefulness of evaluation results (predictive validity). When comparing evaluation methods, it generally does not seem a priori justifiable to consider one method as the “criterion instrument” for the other.

It could be argued that predictive validity is a hotter issue for expert evaluations and methods using self-reporting participants than for methods using behavioral observation. In the case of the first two types of methods, the evaluation results give only indirect indications of reader problems mediated by the judgments of participants. In the case of behavioral observation, it
is, in principle, possible to trap reader problems while they occur. On the other hand, it may not be justified to assume that the scenarios used in a usability test will cover all relevant aspects of document quality. In our view, more systematic research exploring the predictive validity of, say, think-aloud user protocols will be very welcome.

An important factor that has been neglected in the research so far is the initial quality of documents tested. Especially in the case of research into the predictive validity of methods, it is important to know the status of the document used: is it a notoriously problematic document or is it the best possible draft professional designers could produce? In our view, it is not a sensible option to explore the benefits of evaluation methods using documents that are clearly below normal standards of document design, documents that could easily be improved without the help of the evaluation method. For the interpretation of future validation research, it is very important that an indication is given of the status of the document used.

As we have shown, there are various ways of investigating the predictive validity of methods, each with its own strengths and weaknesses. In the most current approach so far, with revision as an intermediate step, it is crucial that the reviser’s input is kept under control. Only then, improvements in the document can be attributed to the evaluation results. Research in which experts assess the severity of the problems detected can be expected to suffer from a general lack of agreement among experts. Research using the error-detection paradigm, with the reader problems of interest inserted in the document by the researcher, can only be used if there is additional evidence that the problems included are real reader problems.

Research into the congruent validity of methods should compare the yield of the methods involved in various respects. The most current type of analysis, the number of problems detected, is a rather ambiguous one. Without indications of problem severity, the number of problems detected is not very informative. Other analyses that may be more informative relate to differences and similarities in the “evaluation standards” of methods (i.e., the amount of attention paid to various types of problems) and the degree of overlap between methods.

On the topic of sample composition, very little research has been carried out so far. The available studies suggest that participants’ background characteristics may affect the feedback collected. More research is needed that systematically explores the influence of participant characteristics on the feedback collected under various methods.

With respect to sample reliability, the Monte Carlo approach seems to be a fruitful “qualitative” alternative for the statistical procedures that can be used in the case of quantitative methods. It may be useful to combine the “added-value” analyses with analyses reflecting the degree of overlap between two subsamples. Caution is needed in immediately adopting the rather favorable Monte Carlo results obtained by Virzi and Nielsen, both suggesting that a sample of five usability participants will suffice, since other studies came to considerably less favorable conclusions.

Very little is known yet about the implementation of troubleshooting evaluation results during the subsequent phase of revision. The scarce and small-scale research cited in this article gives an interesting sample of possible research topics and designs in this area, but offers no (or at best very preliminary) answers to the questions investigated.

With the overview presented in this article, we hope to have demonstrated that there is work to be done in the area of document evaluation and that evaluation methodology is a captivating research object. Both scholars and practitioners may fruitfully contribute to our knowledge about evaluation methods. Scholars may contribute in a programmatic way, focussing a series of studies systematically on methodological aspects of certain evaluation approaches. Practitioners should be encouraged to spend additional research efforts on the evaluation data they have already collected. They might, for instance, compare the feedback collected using a particular method with the feedback from another approach.

**APPENDIX**

**DESCRIPTIOnS OF METHODS MENTIONED**

**Behavior Coding:** The behavior of interviewer and respondent is recorded and coded for the occurrence of specific behavior types which may be indicative for problems in the questionnaire [45], [46], [66].

**Beta Test:** A preliminary version of a computer interface is given to volunteers, who use it in their own environment. The participants describe the problems they experience on paper and fill in a questionnaire about their experiences [68].

**Claims Analysis:** Experts evaluate a computer interface by attaching explicit claims about usefulness and usability to specific stages of realistic task scenarios [59].

**Cognitive Walkthrough:** Experts evaluate a computer interface by following realistic task scenarios and exploring the difficulties users might have during task performance [57]–[59].
Comprehension Test: Participants are asked a number of comprehension questions about a document. This results in an overall impression of the document’s comprehensibility and in the detection of more or less specific areas with comprehension difficulties [18], [20].

Forum Test: A preliminary version of a computer interface is made available via a computer bulletin board. Meanwhile a “forum” is started on a bulletin board, making it possible for participants to discuss the pros and cons of the artifact [68].

Functional Evaluation Checklist: Experts evaluate a document using a checklist based on an analysis of functions that the document has to fulfill [56].

GOMS Analysis: Experts evaluate a computer interface by analyzing which methods users can choose to achieve their goals with it and, in the case of alternative methods, how they may select the methods depending on the context [59].

Heuristic Evaluation: Experts evaluate an artifact (document or interface) using a set of guidelines or principles for effective design [25], [35], [57], [59].

In-Text Feedback Questions: Participants are asked to use a document independently and answer written feedback questions that are inserted following textual units [22].

Interviewer Debriefing: A questionnaire is evaluated by interviewers who try out the questionnaire with potential respondents and reflect on the problems encountered [45], [66].

Logged Data: Participants are asked to carry out tasks using a computer interface. Their sequence of actions is recorded in a log file [64].

One-to-One Evaluation: Individual learners are observed while using instructional materials. The evaluator may interact with the learners about their experiences and their suggestions for improvement [21], [25]–[27], [33], [65].

Plus–Minus Method: Participants are asked to read a document and put pluses and minuses in the margin for positive and negative reading experiences. After that, the reasons for the pluses and minuses are explored in an individual interview [19], [34], [41], [47]–[51], [56], [76].

Probing Questions: A questionnaire is investigated using additional questions exploring, for instance, the way participants interpreted the questions, the way they retrieved the information required, and the clarity and suitability of response categories offered [46].

Retrospective Think-Aloud Protocols: Participants are asked to carry out tasks using an artifact (document or interface). After that, a video recording of their task performance is shown, and they are asked to think aloud as if they were carrying out the task. The data used are the behavioral observations and the retrospective think-aloud protocols [64].

Signaled Stopping Technique: Participants are asked to record their own stopping behavior while reading a document by writing down a signal in the margin. After that, the reasons for the reading interruptions are explored in an individual interview [24].

Small-Group Evaluation: A group of learners is exposed to instructional materials under normal circumstances of use. Various kinds of data may be collected: test scores, observations, written comments, and evaluative responses [21], [26], [27], [65].

Think-Aloud Reader Protocols: Participants are asked to read a document and think aloud. The think-aloud data may capture the participants’ process of understanding the document and their evaluative responses [54].

Think-Aloud User Protocols: Participants are asked to carry out tasks using an artifact (document or interface) and to think aloud during the performance of the tasks. The data used are the behavioral observations and the think-aloud protocols [28]–[32], [37], [50], [54], [68], [69].

User Action Notation: Experts evaluate a computer interface by specifying the sequence of actions users have to take in order to carry out realistic tasks [59].

REFERENCES


Menno de Jong is an associate professor of Communication Studies at the University of Twente, Enschede (The Netherlands). His research interests concern the use and methodology of applied research to optimize communication. He has published about formative text evaluation, Web evaluation, usability, and document design. His Ph.D. Dissertation (1998) deals with the value of the plus-minus method for pretesting public information brochures.

Peter Jan Schellens is a professor of Communication Studies, especially linguistic aspects of communication, at the University of Twente, Enschede (The Netherlands). His research interests include document design, text evaluation, and argumentation analysis. He coordinates a research program on verbal communication and document design. He is scientific director of the Twente Institute of Communication Research (TwICoR).