search cannot always be justifiably evaluated using a given complex search characteristic. We will refer to those situations where it can be done as suitable for a complex search characteristic considered, and those situations where it is impossible we will refer to as nonsuitable. With this statement in mind, it is easy to imagine that in the evaluation of a certain macroevaluated object, the majority of values of a complex search characteristic used were obtained in nonsuitable situations.

It is clear that the evaluation based on a value derived from averaging such nonsuitable values is not very reliable. Here is a particular example. Suppose that we want to select the best search strategy from two strategies and for this purpose we have performed a search using the same search requests. Further, assume that the search using the first strategy essentially resulted in outputs that contain exactly 1 document and that it is pertinent, with 20 pertinent documents available in the collection of documents. (Then, the complex search characteristic I_1 has achieved values equal to 1.05.) Finally, assume that the search using the second strategy essentially resulted in outputs containing 10 pertinent and 10 nonpertinent documents, with 20 pertinent documents available in the collection of documents (the value of the complex search characteristic I_1 is 1). If two search strategies are evaluated on the basis of the values derived as the result of averaging values of characteristic I_1, then the first strategy is more likely to be preferred over the second strategy, whereas in a real situation for the vast majority of search requests, the search using the second strategy can provide higher functional effectiveness in comparison with the search using the first strategy. Thus, the evaluation based on a value derived by averaging values of a certain complex search characteristic cannot necessarily be trusted.

But in which cases can such an evaluation be trusted? Obviously, when all averaged values of a complex search characteristic used were obtained in suitable situations. However, from a practical point of view, such a requirement seems to be too strong. It would therefore be more appropriate to introduce into consideration a value representing a ratio (expressed as a percentage) between a number of those values of a complex search characteristic that was obtained under suitable situations and a number of all found values of this characteristic, and then to determine in test runs the threshold of this value upon attainment of which the evaluation discussed can be trusted. It should be pointed out that in this case the evaluation will involve averaging not all available values of a complex search characteristic but only those that were obtained in suitable situations. The threshold just mentioned will be called the "confidence threshold." It is clear that, in using a specific complex search characteristic, the wider the circle of suitable situations for this complex search characteristic, the more frequently the confidence threshold will be achieved. So not all complex search characteristics are equal from the point of view of their use in the evaluation process. Naturally, the wider the circle of suitable situations for a particular complex search characteristic, the more appropriate its use in an evaluation.

Analysis of the domains of applicability for complex search characteristics I_1 and I_2 (these domains were discussed in Chapter 10) shows that these domains are "wide" enough (and, correspondingly, a circle of suitable situations for complex search characteristics I_1 and I_2 is rather wide) and implies that it is appropriate to use these characteristics in the discussed evaluation. Also, the domain of applicability of characteristic I_2 turned out to be "wider" than that of the characteristic I_1. Furthermore, this domain was found to be so "wide" that the following statement important to information practice was considered in Chapter 10 as a rather moderate assumption: If the precision level of a search is not less than 0.5, its functional effectiveness may be justifiably evaluated on the basis of complex search characteristic I_2 practically at any attained value of the given characteristic; that is, virtually any situation would be suitable for this complex search characteristic.

The requirement of attaining a precision of no less than 0.5 appears reasonable in modern practice. Thus, taking into account the preceding statement, as well as the fact that the confidence threshold would more likely be lower than 100%, we can justifiably claim the following: If in the discussed evaluation complex search characteristic I_2 is used, then the result of such an evaluation can be practically always trusted. By the way, if nevertheless a need arises to confirm that the confidence threshold in evaluation has really been achieved, in the case of characteristic I_2 this should not cause any difficulties. Indeed, in this case making a decision on whether the situation considered is suitable can be justifiably based only on the analysis of an achieved precision level. To clarify, in this case whether or not the confidence threshold has been attained in the evaluation, there is no need to determine in the analyzed situations the recall level. On the other hand, to resolve the same problem in the case of complex search characteristic I_1, it is necessary to determine the recall level. (This statement follows from the results shown in Chapter 10.) Further, it would seem useful to consider the suitability of applying other complex search characteristics in the discussed evaluation, but we cannot give any recommendations about applications of these characteristics because we do not know of any studies where these characteristics were analyzed in the same manner as that presented in Chapter 10 regarding characteristics I_1 and I_2.

In summary, recall that the functional effectiveness of a document search may be evaluated on the basis of not only complex search characteristics but also on general search characteristics or, more precisely, their combination. In this connection there have been attempts to evaluate macroevaluated objects also using a set of averaged values of such characteristics. It should be acknowledged that the evaluation of macroevaluated objects based on a set of averaged recall values and averaged precision values is very popular in information science. Nevertheless, such an evaluation cannot be trusted in all cases. Consider the following example. Suppose that we evaluate a certain retrieval service within the scope of which a search using a number of search requests needed for performing the