the type of question and degree of comparability of the questions between the instruments. In addition to measures of agreement, we present an assessment of bias—i.e., whether one instrument consistently rated the indicator higher (or lower) than did the other instrument.

The first indicators are objective measures of the provider’s actions with the client. Agreement was fair to good (kappas, 0.57–0.71) on three of these indicators and poor on the fourth (0.30). The only indicator for which we found evidence of bias was whether the return visit was discussed. This finding reflects that in Zimbabwe, clients greatly underreported the occurrence of such discussions.

A second set of indicators measure the information exchange that occurred between the client and provider on different topics; we also considered these to be objectively measured. All but one of these indicators had fair to good agreement (kappas, 0.47–0.69). As we noted earlier, clients frequently reported receiving information not recorded during the observation, probably because this occurred during an unobserved part of the visit.

This set of indicators includes one that was calculated from questions that were less than comparable on the two instruments—whether the provider gave the client accurate information on how to use the method she chose. On this indicator, we saw some evidence of better results from the exit interview, primarily because often in Uganda and Zimbabwe, new clients correctly reported key information on how to use their chosen method, yet observers did not record that accurate information was provided. While a client’s knowledge of her method may have been obtained during the visit with the provider, in many cases, she may have already had correct information or obtained it from other sources at the health facility.

The third set of indicators measure interpersonal involvement; we deemed these to involve more subjectivity. Surprisingly, agreement between results from observations and exit interviews was excellent for all of these (kappas, 0.75–0.98), and was actually higher than for the more objective indicators. The two indicators that assessed whether privacy was adequate revealed bias: Fewer clients than observers reported adequate privacy for counseling or examinations. This difference may suggest that observers’ perceptions of what constitutes privacy differ from clients’, possibly because of observers’ familiarity with the health care system and its norms. Effective training of observers can reduce interrater reliability, but it cannot eliminate this difference in perception. For the remaining two indicators, the responses showed no evidence of bias.

Kappa coefficients for the 14 indicators for the three countries combined are presented in Figure 2. Agreement ranged from poor to excellent. Both kappa coefficients and percentage agreement (which ranged from 63% to 99%—see Table 2) gave very similar findings for the indicators.

**Discussion**

Overall, the results obtained from observations and client exit interviews were highly comparable for most indicators. To the extent that discrepancies occurred, the major reason for these discrepancies was that clients received information from sources other than the observed client-provider interaction. Such other sources as group talks and supplemental counseling sessions need to be taken into consideration in interpreting the results of this study and in using these instruments in the future.

The consistently high ratings for indicators measuring interpersonal relations may reflect that providers were on their best behavior, since they knew they were under study (the Hawthorne effect). This upward bias, however, should have affected the responses from observations and exit interviews equally—i.e., observers would have recorded better behavior on the part of providers, and clients would have reported the same during exit interviews. (Whether the high ratings on these subjective measures are due to the Hawthorne effect cannot be addressed with the current data.) Moreover, despite the presence of observers, many objective indicators suggest serious deficits in quality in many areas.

A major drawback of exit interviews is courtesy bias. One would expect that indicators measuring subjective states such as attitude, opinions or feelings would be more susceptible to courtesy bias than more objective measures. We did not find this to be true in our data. In fact, agreement was highest on the indicators that we considered more subjective. It is possible, however, that providers were on their best behavior because of the observer’s presence, and clients were truthfully reporting good interpersonal relations. Whether the clients would have been as truthful in a situation where providers were rude or unresponsive is not known.

Other sources of error are also possible. Recall bias may account for a client’s “forgetting” that a specific instruction or particular information was provided during the visit. Given that the client was interviewed immediately following the visit, she may have had time to think about the session and process all of the information that she received. This can be seen with indicators that are measured similarly in both instruments: Whether the provider discussed the return visit with the client is an example. Though the questions are relatively straightforward, clients sometimes reported that this was discussed while it was not, and vice versa.

We also considered whether interviewee fatigue and a desire to terminate the interview quickly may have introduced two