Scientific Reports

NUMBER 54 APRIL 1984

THOMAS W. PULLUM
NURI OZSEVER
TRUDY HARPHAM

An Assessment of the Machine Editing Policies of the World Fertility Survey

INTERNATIONAL STATISTICAL INSTITUTE
Permanent Office. Director: E. Lunenberg
428 Prinses Beatrixlaan, PO Box 950
2270 AZ Voorburg
Netherlands

WORLD FERTILITY SURVEY
Project Director: Halvor Gille
35–37 Grosvenor Gardens
London SW1W 0BS
United Kingdom

The World Fertility Survey is an international research programme whose purpose is to assess the current state of human fertility throughout the world. This is being done principally through promoting and supporting nationally representative, internationally comparable, and scientifically designed and conducted sample surveys of fertility behaviour in as many countries as possible.
The WFS is being undertaken, with the collaboration of the United Nations, by the International Statistical Institute in cooperation with the International Union for the Scientific Study of Population. Financial support is provided principally by the United Nations Fund for Population Activities and the United States Agency for International Development.
This publication is part of the WFS Publications Programme which includes the WFS Basic Documentation, Occasional Papers and auxiliary publications. For further information on the WFS, write to the Information Office, International Statistical Institute, 428 Prinses Beatrixlaan, Voorburg, The Hague, Netherlands.

L'Enquête Mondiale sur la Fécondité (EMF) est un programme international de recherche dont le but est d'évaluer l'état actuel de la fécondité humaine dans le monde. Afin d'atteindre cet objectif, des enquêtes par sondage sur la fécondité sont mises en oeuvre et financées dans le plus grand nombre de pays possible, Ces études, élaborées et réalisées de façon scientifique, fournissent des données représentatives au niveau national et comparables au niveau international.
L'Institut International de Statistique avec l'appui des Nations Unies, a été chargé de la réalisation de ce projet en collaboration avec l'Union Internationale pour l'Etude Scientifique de la Population. Le financement est principalement assuré par le Fonds des Nations Unies pour les Activités en matière de Population et l'Agence pour le Développement International des Etats-Unis.
Cette publication fait partie du programme de publications de l'EMF, qui comprend la Documentation de base, les Documents Non-Périodiques et des publications auxiliaires. Pour tout renseignement complémentaire, s'adresser au Bureau d'Information, Institut International de Statistique, 428 Prinses Beatrixlaan, Voorburg, La Haye, Pays-Bas.

La Encuesta Mundial de Fecundidad (EMF) es un programa internacional de investigación cuyo propósito es determinar el estado actual de la fecundidad humana en el mundo. Para lograr este objetivo, se están promoviendo y financiando encuestas de fecundidad por muestreo en el mayor número posible de países. Estas encuestas son diseñadas y realizadas científicamente, nacionalmente representativas y comparables a nivel internacional.
El proyecto está a cargo del Instituto Internacional de Estadística en cooperación con la Unión Internacional para el Estudio Científico de la Población y con la colaboración de las Naciones Unidas. Es financiado principalmente por el Fondo de las Naciones Unidas para Actividades de Población y por la Agencia para el Desarrollo Internacional de los Estados Unidos.
Esta publicación ha sido editada por el Programa de Publicaciones de la EMF, el que incluye Documentación Básica, Publicaciones Ocasionales y publicaciones auxiliares. Puede obtenerse mayor información sobre la EMF escribiendo a la Oficina de Información, Instituto Internacional de Estadística, 428 Prinses Beatrixlaan, Voorburg-La Haya, Países Bajos.

Scientific Reports

An Assessment of the Machine Editing Policies of the World Fertility Survey

THOMAS W. PULLUM
University of Washington and WFS Consultant

NURI OZSEVER
WFS Central Staff

TRUDY HARPHAM
WFS Central Staff

The recommended citation for this publication is:

Pullum, Thomas W., Nuri Ozsever and Trudy Harpham (1984). An Assessment of the Machine Editing Policies of the World Fertility Survey. *WFS Scientific Reports* no 54. Voorburg, Netherlands: International Statistical Institute.

Printed in Great Britain
by Spottiswoode Ballantyne Limited, Colchester and London

Contents

Preface

A critical assessment of survey experience and data quality is an integral part of the WFS programme. This assessment aims at ensuring that analyses are carried out with as full an understanding as possible of the quality and reliability of the data and at drawing lessons for the better conduct of future surveys.

One of the many ways in which WFS attempted to ensure that its surveys were of the highest possible quality was by editing the data for internal consistency. This paper focuses on the type of editing which takes place between data entry and the construction of the first Standard Recode File. Initially considered to be an essential but rather routine activity, it was in fact the source of major delays in the publication of survey results in many participating countries. This assessment describes the procedures used, estimates their cost, and tries to account for the delays. More importantly, it tries to determine whether elaborate machine editing actually affects analyses and interpretations. The conclusions should be very useful for future surveys carried out under similar circumstances.

Halvor Gille
Project Director

1 Problems and Objectives

From the very inception of the World Fertility Survey, one of its major ambitions has been to produce data of first-rate quality. This goal has been sought in every phase of operation, including questionnaire design, sample selection, interviewing, data processing, and analysis of findings. The pursuit of quality has necessarily incurred some delays and other costs in each of these phases. This paper, part of a general assessment of WFS activities, focuses on the various costs and benefits attached to the machine editing of data which accompanies the preparation of standard computer files for each country.

Although this report is based exclusively on WFS experience, and functions in part as a summary and appraisal of that experience, it is also intended to have some value for the planning of similar surveys in the future. WFS surveys have, of course, a large number of shared characteristics, which have permitted some standardization of editing procedures. Other multi-survey programmes would have similar economies of scale, and would confront similar kinds of editing costs and benefits. Single *ad hoc* surveys do not present such opportunities for sharing the costs of software development, for example, but certainly the benefits of editing should be comparable. We hope to develop some criteria and decision-making rules which could include a wide range of substantive research topics and survey designs.

The fundamental questions are these: what is the purpose of editing survey data; why, in particular, did WFS adopt such stringent editing policies; and were these reasons justified? As indicated above, WFS believed that editing would improve the quality of the final product. Existing documentation does not spell out the link between editing and data quality nor precisely define data quality, but we believe that at least the following four assumptions led to the present policy. First, editing is believed to produce a gain in the yield of the fieldwork. Major components of the cost of a survey, obviously, are the design of the questionnaire and sample and the interviewing of respondents. Editing, at a marginal cost which is small compared to these, will make more cases and specific items usable. It will minimize the number of responses which must be excluded from analysis. Secondly, editing is believed to improve the validity of the findings. That is, estimates based on edited data will tend to be closer to the population values which they are intended to estimate. This belief is based on the supposition that discrepancies in the data tend to be systematic rather than random, and introduce certain kinds of bias. Thirdly, editing improves the correspondence between the structure of the sample and of the questionnaire on the one hand, and the structure of the responses on the other. Internally consistent data greatly facilitate tabulation and other forms of analysis, even though the conclusions may not be affected. Internal consistency is also believed to increase the user's sense of confidence in the data.

The fourth main reason why WFS data were edited so carefully was the perception that this practice was a hallmark of professional survey research. One may suggest that this was in fact the fundamental justification. Machine editing of data from fertility, KAP, or demographic surveys in developing countries was not standard practice (although field editing has a long history) much prior to the first WFS surveys in 1974, partly because of the lack of adequate computing facilities in such countries. However, machine editing had become a routine phase in the data processing of major social surveys in the United States, at least, during the 1960s. When WFS policies were being formulated, it was accepted with very little question that this practice was one of the essential ingredients of sound methodology and that WFS surveys should in fact serve as a vehicle to introduce modern editing to statistical offices in developing countries.

In this report we shall step back a bit from the assumptions or beliefs described above, and shall reconsider in retrospect whether these policies were justified. At issue is whether, beyond some point, the measurable returns from continued refinement of the data base are so small that it is inefficient to make further improvements. This assessment does not question the basic desirability of high quality; it simply allows for the possibility that some apparent increments in quality may absorb time and resources which could be better spent elsewhere. To the limited extent that relevant data are available for this purpose, we shall attempt to reconstruct costs and to simulate alternative policies and to reach some conclusions about the cost-effectiveness of WFS policies.

This phase of the general assessment has been partially motivated by conflicting reactions from users of WFS data. On the one hand, some users have been greatly pleased with the fact that the data tapes prepared by WFS —in particular, the Standard Recode Files—are relatively free from internal inconsistencies. These files are written in a standard format and are accompanied by thorough documentation, so that users can proceed rapidly to second stage and comparative analyses. But on the other hand, some other users are frustrated by the length of time between the fieldwork and the emergence of the principal results, some of which delay is attributable to data editing. The elapsed time from the completion of fieldwork until the completion of the First Country Report—to say nothing of its publication—is sometimes more than three years and rarely less than two. Claims have been made of six to twelve months of unnecessary delay due to editing. If these are correct, then a change in editing policies might produce a marked reduction in the total elapsed time. It has also been argued that extensive editing puts undue

distance between the analyst and the primary data and obscures an evaluation of data reliability.

Perhaps the key to the difference between these two perceptions of WFS data lies in the attitude toward the First Country Report, which is not prepared until the editing has been completed. The format for this report and its tabulation plan were developed under certain basic assumptions. The first of these was that many countries would never be able to proceed beyond the level of the First Report because of limited research personnel, facilities, and funding. It was therefore believed that the First Country Report—so named out of the hope, rather than the expectation, of later reports—should have a high level of completeness and accuracy. As it included a large set of detailed tables, at least some further research would be possible from those tables by other analysts. Secondly, it was assumed that the preparation of an advance report, say, with data that were less thoroughly edited and on a restricted set of variables would draw resources away from the First Country Report and delay its completion even further. And thirdly, it was believed that WFS would suffer a loss of confidence if data were released which showed inconsistencies. It would be embarrassing (it was believed) to produce tables in which corresponding totals or subtotals did not agree exactly, or to revise the estimates of important rates, means, or proportions in later publications. Perhaps these assumptions now need to be modified on the basis of experience.

It will help to place data editing in its proper context if certain issues are raised at this early point.

The primary achievement of editing or cleaning is to detect whether the various responses are consistent with one another and with the basic format of the survey instrument, and to resolve any detected inconsistencies through adjustment. Editing is not properly described as the correction of errors; conversely, a good many errors, of many kinds, will not even be touched by the editing process. The point here is that there is no way of genuinely validating any of the responses.

Errors or inconsistencies can arise at several phases. There is often some ambiguity about their origins, but a rough classification is as follows:

1 Response error: because of misunderstanding of a question, recall failure, or attitudinal ambivalence, a respondent may supply invalid or unreliable information.

2 Interviewer error: the interviewer may incorrectly record the information, whether because of a communication failure with the respondent, because of a misunderstanding of the survey instrument or because of carelessness.

3 Coding error: responses may be incorrectly translated into numerical codes.

4 Data entry error: keypunchers may read the codes inaccurately, reverse digits, shift the codes to incorrect locations, and so on.

5 Programming errors: any program which edits, recodes, or constructs files has the possibility of altering the data in an unintended manner.

6 Specification errors: although these are sometimes classified with programming errors, they are often attributable to an analyst rather than to a programmer or computer as such. For example, a researcher may design a complex summary variable without in fact having categories which are complete and mutually exclusive. The design may be accurately implemented but logically defective.

The possibility that an error can be detected and corrected varies roughly in relation to the stage at which it occurred. Machine editing is obviously unable to detect all of the problems which the file may actually possess. For example, many of the background and attitudinal variables will be flagged only if a code is out of range. Obviously, other errors can occur but remain within the legal range. Regarding consistency, it is possible for an erroneous response to be consistent (ie not inconsistent) with other responses whether those are themselves erroneous or not. Even among the dates and intervals, which are checked carefully, there is necessarily a threshold level for each check. The most that can be said is that errors are more likely to be detected if they exceed the threshold. For example, the 'date edit' compares date of birth and date of marriage and prints out an error message if the difference—the age at marriage—is calculated to be less than some threshold level such as 12 years. Such cases will then be reviewed. However, a variety of data entry or reporting errors is possible, involving date of birth and date of marriage and related variables, which will not be signalled by the above check nor any similar ones. The subject of response errors in the birth histories has received a great deal of attention, and it is generally agreed that editing and imputation have only a limited impact on them.

In addition, of course, the estimates derived from a survey are subject to some degree of sampling error. WFS surveys are usually large, but some of the important estimates—such as the age-specific fertility rates and the proportions of women using specific contraceptive methods—are nevertheless subject to considerable error (Little 1982). Fortunately for present purposes, WFS has carefully prepared estimates of the sampling error of important quantities.

When viewed in the context of sampling error, response error, data entry error, and other non-sampling error, the significance of sophisticated editing procedures may appear somewhat diminished (O'Muircheartaigh and Marckwardt 1981). One might even urge that a perfectionist policy is superficial and deceptive, in that it may give the user a false sense of confidence in the data. This is at least a viewpoint which should be recognized and given an even-handed consideration. It is difficult, however, to assess the improvement in data quality from editing relative to the total survey error, because we do not know the magnitude of the total survey error.

These comments lead to some subsidiary issues. One of these is bias. Is it possible to say that the edited responses are less biased than the unedited ones? In one sense this is

probably true. Non-response and some kinds of detectable inconsistencies are more common in some subpopulations, such as the poorly educated, than in others. If estimates are made for the total sample, and these subpopulations are differentially under-represented, then a bias is clearly introduced into the estimates. For example, an overall mean is a weighted average of subgroup means, and if a subgroup has a high level of non-response then it will be given too little weight in the overall mean. However, within homogeneous subgroups it is less clear whether the edited or unedited estimate is more accurate, except for the correction of obvious data entry errors. In the empirical investigation it will be seen that the two estimates may differ, but usually by a small amount, and one cannot be certain that the edited estimate is closer to the population value, even though of course we would like it to be.

2 A Description of the Current Machine Editing Procedures

Two types of editing routinely precede data entry in WFS surveys. These are the field and office edits, which were typically the only kinds of editing possible for demographic and fertility surveys prior to the computer sophistication of the 1970s. They are not the subject of the present assessment, and are described in detail in WFS core documentation. In brief, field editing is the responsibility of each interviewer and his or her supervisor, and is intended to detect inconsistencies recorded on the questionnaire while there remains the possibility of going back to the respondent for a resolution. This is probably the most important phase of editing in the sense that there is still access to the respondent. Office editing is usually carried out in conjunction with coding, and involves a check of the numerical computer codes as well as the responses recorded by the interviewer. Obviously, the field and office checks cannot be as numerous or as complex as the computer checks, but they are carefully designed to detect the inconsistencies with the most serious ramifications. This report is based on the assumption that such editing has been done, and any inferences drawn from our analyses must be restricted to the marginal role of machine editing.

As described in the 'Data Processing Guidelines', there are three principal phases in the cleaning of the raw data file from each country. The first is the 'structural edit'. During this phase, the major concerns are the uniqueness and proper sequencing of identification codes; the existence of the necessary cards for each respondent; and the proper sequencing of those cards. If such things are not checked, then it is difficult to interpret the file properly. It will be assumed that the structural edit must always be done and it is not a part of the present assessment. It may be observed that there is only a fine line of distinction between some structure checks and some of the subsequent consistency checks, and it typically happens that during the later checks some cards (records) or entire cases will be deleted because of duplications or identification errors or large discrepancies in the birth histories which might more properly be labelled as structural errors.

The second phase of the process is the 'consistency edit', which includes range, skip, filter, and other consistency checks on variables other than dates, which are deferred to the next phase. These different kinds of checks are made with their own programs: the range checks; the skip, filter, and table (ie birth and marriage history) checks; and miscellaneous consistency checks. Because each country has a somewhat different questionnaire and raw data file format, these programs have different specifications in each country. A special language, CONCOR, is generally used, but it requires some quite detailed programming of the specific edit checks. The 'consistency edit', like the 'structural edit', is virtually always conducted in the country itself, using available hardware. CONCOR is usually not available, and must be installed and tested. The edit specifications must be written with the heavy involvement of WFS data processing staff from London. Thus, there is substantial potential for delay even before the first of the consistency edit checks are produced, just from the preparatory work required. If all goes well, of course, then this advance work will have been completed before the end of the structural edit, and the consistency checks can begin straightaway.

The third phase of editing is the 'date edit', involving consistency checks among the birth and marriage histories and other reported dates and intervals on such events as child deaths, breastfeeding, sterilization, etc. This includes even such checks as agreement between the reported number of sons ever born and the number of male births in the birth history. This phase begins with the preparation of a specially written date extraction program, which rearranges all of the relevant dates and intervals for each respondent on to a single record with a standard format. The data file resulting from this operation is then submitted to the date edit program. This program is known as DEIR (for date editing, imputation, and recoding). Up to a point, the date edit proceeds similarly to the consistency edit, with inconsistencies resolved so far as possible by references to the questionnaires. At this point, the final raw data file is produced. This file has the same format as the earlier raw data files, which varies from one country to another; the changes resulting from the date edit runs are made on this file.

Apart from exceptional cases, the conventional editing as such ends with the construction of the final raw data file. When this file is ready, the Standard Recode File is then constructed with the DEIR package and a specially written recode program. The DEIR package accomplishes all of the following in one automatic step: (1) it reformats the individual respondent's extracted data on to a single long record, using a standard format for all countries; (2) it adjusts and imputes missing or incomplete dates and intervals, sometimes making changes which were not flagged earlier, in order to improve consistency; and (3) it constructs a number of date-dependent standard recode variables which are called for in the tabulation plan. The recode program is then used to construct additional standard recode variables. The resulting Standard Recode File is the basis for virtually all analysis, beginning with the tables for the First Country Report.

In each of the three major stages of editing just described, three steps are cycled through iteratively. These may be labelled 'checking', 'reconciliation', and 'updating'. We shall use these terms in preference to others which suggest that editing consists of error correction; only a subset of all editing is properly described in this way. Checking and updating are wholly computerized operations, applied

to data in computer files. The checking is performed with a computer program, and a so-called error statement is printed out for each record which fails the checks. The error printout includes some diagnostic remarks and lists out the codes for the variables involved.

The reconciliation step is performed manually except for the final date imputation. That is, a knowledgeable individual examines the original questionnaire alongside the error printout and writes out a correction statement. Sometimes there is obvious evidence of a data entry error; sometimes an inconsistency occurred in the field but was not caught at that level, and an informed guess at the true situation must be made. For example, a filter may be inconsistent with the items on which it is supposedly based; and sometimes there is no choice during the reconciliation step but to employ a 'not stated' or 'no response' default if consistency is to be achieved. Such a default leads to harmony of a sort but through the partial loss of a case.

The reconciliation is followed by a case update. In the typical batch-processing mode, a large number of these will be read in one run, with the corrections entered on punch cards. Because there is a fairly high incidence of keypunch errors in the updating step, and because of the multiple interdependencies of responses, the checking program is rerun after each set of updates. These steps are repeated in sequence until no edit violations are found, at which point the next step in the editing is taken up.

In the second part of the DEIR package, the date imputation procedure may be regarded as a fully automatic combination of the checking, reconciliation, and updating steps. Here, changes are made without reference to the original questionnaire and without any human participation—although, of course, many assumptions have been built into the program and through the adjustment of parameters it is possible to tighten or to relax some of these assumptions.

After the first Standard Recode File has been constructed in the country, some modifications are still possible. Some countries have had four or more versions of this file. Changes at this point typically involve the addition of new recoded or country-specific variables, the correction of earlier programming errors, the resolution of minor inconsistencies which were inadvertently omitted in the earlier specifications, and so on. These updates are done in London. Such revisions are not associated with delays in the issuing of results and will not be discussed further in this report.

The WFS approach to editing may be described as a perfectionist policy. Three possible alternatives to this policy will now be described briefly. In each of them, all of the regular edit checks would be made, using an edit program prepared well in advance of its use. But in each alternative, all inconsistencies which remain after some preset level would be resolved by an automatic default, such as the use of 'not stated' codes.

The first alternative may be labelled a cost-specific policy. Its fundamental difference from what is currently done would be that the allowable budget and/or elapsed time for editing would be preset at some level, such as an elapsed time of three months, and after that point the editing would terminate. Since there are several phases to the editing, this change would require subsidiary cutoff levels for each phase.

Secondly, a criterion for termination which is defined in terms of potential benefits rather than costs leads to a benefit-specific policy. For example, if fewer than five per cent of the cases have an inconsistency involving an important variable, then one might decide to do no further editing.

The third option may be described as a purpose-specific policy. At present, all data analysis is postponed until the Standard Recode File has been produced, that is, until virtually all editing and date imputation have been completed. The only editing done later is an occasional minor update of this file. If it were found that, beyond a certain point, editing would continue to have value for second stage analysis but was pointless for the First Country Report tables, then one could attempt a better co-ordination between the editing and the stage of analysis. Specifically, one could edit up to a specified threshold; produce the First Country Report tables; and then resume the editing for more sophisticated purposes. The basic difference from the perfectionist policy would be that at some point the editing would be interrupted and the basic tables produced so that the First Country Report could appear sooner.

One could also consider, of course, combinations of these three alternative strategies. More possibilities would be opened up if alternative phases of analysis were considered—for example, an advanced version of the First Country Report, or some other replacement of that Report by a series of smaller reports. The range of such alternatives is almost without limit.

3 The Costs of Machine Editing

In order to establish the need for the present assessment and the consideration of alternative policies, we shall now estimate the cost of machine editing incurred by WFS. The remainder of this report will then estimate the value of the editing for the analyses, in order to determine whether the costs were justified.

The cost of editing is measured by its absorption of resources which could be used in other ways. The two principal aspects are the monetary cost and the delay in general accessibility of the data. The monetary cost includes person-days of local or in-country staff time; person-days of WFS professional staff time, including a portion of travel costs if travel to the country is required; computer time; and lesser miscellaneous costs.

The direct costs during the interval which could be attributed to editing—for example, computer time and staff time—are simply not available to us. However, the average for a country will certainly be at a rate of tens of thousands of dollars per year. Much of this time is spent within the participating country, rather than in London, with a cost which may be lower in dollars but high in terms of tying up skilled research and programming staff for long periods.

Elapsed time is probably the most important and most conspicuous cost of the editing, but it is also difficult to assign a monetary equivalent. We shall now propose a simple conversion.

Delays in issuing survey results seriously impair the usefulness of those results. This is clearly true for policy and programmatic uses, and appears also to be true for some academic or scholarly uses as well, possibly apart from comparative, methodological, and historical research. We shall therefore attempt to convert elapsed time to a monetary equivalent through the notion of depreciated value. Considering the dates of analyses of WFS surveys in relation to their dates of fieldwork, it appears that relatively little interest remains in a survey after about five years. One notes also that several countries follow a quinquennial schedule of population censuses, and several others aspire to that level of frequency. In the Philippines, national demographic surveys are conducted on a five-year basis; the 1978 Republic of the Philippines Fertility Survey was third in this series and a fourth one occurred in 1983. Five years is the typical length of age intervals, forecasting intervals, and national planning cycles. Following these admittedly somewhat arbitrary conventions, we suggest that a linear five-year depreciation schedule be applied to a WFS survey, such that after five years a survey has altogether lost its value for current purposes. We shall also assume, arbitrarily, that 60 per cent of the value of the survey is for current purposes, and 40 per cent is for long-term purposes and is not reduced by delays of any kind at any point. Therefore, it is suggested that every month of delay during the first five years reduces the survey's value by one per cent of its total cost. The so-called total cost will be taken to be the sum of all expenditures on the survey up to the publication of the First Country Report. For example, if a survey is calculated to cost $500,000, then the pro-rated depreciation in one month will be $5000. It could be argued that the depreciation calculation should not begin at the time of the survey (ie at the date when fieldwork is completed), but rather at the earliest possible date when the report could have been released. However, in terms of policy and programmatic value, the utility of the results begins to decline as soon as the interviewing has been completed, and, if anything, declines most rapidly during the first few elapsed months. One would argue for other refinements of the depreciation formula, but we doubt that they would materially alter the conclusions.

Table 1 shows the number of months elapsed between the end of the data entry and the production of the first recode tape for the WFS surveys. The average interval is 16 months. The interval is an exaggeration of the time devoted to data editing as such and also includes time spent on structural editing, but intermediate dates which might be more appropriate are unavailable. However, we do not believe that the exaggeration is serious. The data processing tasks of running the extraction program and DEIR, both of which should be ready before the start of the interval, are the only substantial tasks other than editing which are required during the interval. Out of the 40 countries for which both dates were available, four required intervals of only four months or less. If the editing were abbreviated drastically, then it would be possible to prepare the Standard Recode File routinely in such a period. We therefore suggest that the average interval and thus the average time to the First Country Report were both prolonged by one full year by machine editing.

We would prefer to provide country-specific estimates and to break them down by levels of editing. The appendix of this report provides an outline for a monitoring scheme which would permit such refined calculations. However, the countries listed in table 1 had a median total survey cost of $590,000 (this includes the survey budget and technical assistance costs). If we assume, as suggested above, that every month of delay reduces the survey's value by one per cent of its total cost, it can be calculated that the average cost of data editing is at least $75,000, including a combination of depreciation and some direct costs. If this average is applied to the 40 participating countries, then a total of at least $3 million of the total WFS budget will have gone to range, skip, and consistency machine editing. It remains to be seen whether an expenditure of this approximate magnitude was justified.

Table 1 Elapsed time between end of the data entry (preparation of first raw data file) and construction of first Standard Recode File

Country	Completion of data entry	Completion of first recode file[a]	Elapsed time in months
Bangladesh	June 76	Nov 77	17
Benin	Sept 82	Aug 83	11
Cameroon	31 Mar 79	Jan 82	34
Colombia	17 Oct 76	Jun 77	8
Costa Rica	Dec 76	Jan 78	13
Dominican Rep.	24 Oct 75	May 76	7
Ecuador	May 80	Mar 82	22
Egypt	Oct 80	Apr 82	19
Fiji	Aug 74	Jun 76	22
Ghana	9 May 80	Nov 81	18
Guyana	10 Oct 75	Nov 78	37
Haiti	20 Feb 78	Jun 80	28
Indonesia	Oct 76	Jan 78	15
Ivory Coast	Aug 81	Mar 83	19
Jamaica	Apr 76	Feb 79	34
Jordan	14 Sep 76	Jul 78	22
Kenya	Oct 78	Jul 79	9
Korea	19 Aug 75	Feb 77	18
Lesotho	Dec 78	Nov 79	11
Malaysia	May 75	May 76	12
Mauritania	Mar 82	May 83	14
Mexico	6 Apr 77	Aug 78	16
Morocco	Feb 81	Jul 83	29
Nepal	30 Sep 76	Nov 76	2
Nigeria	Data entry not complete (Sept 1983)		
Pakistan	Dec 75	Jun 76	6
Panama	May 76	Jan 77	8
Paraguay	Sep 79	Jan 80	4
Peru	Jun 78	Jul 78	1
Philippines	31 Aug 78	May 79	9
Portugal	30 Oct 80	Nov 81	13
Senegal	Feb 79	Jul 80	17
Sri Lanka	Apr 76	Jun 77	14
Sudan (North)	Jan 80	Mar 81	14
Syria	Apr 79	Jan 81	21
Thailand	Nov 75	Mar 76	4
Trinidad and Tobago	Oct 77	Nov 79	25
Tunisia	28 Feb 79	Oct 81	32
Turkey	Feb 79	Mar 80	13
Venezuela	12 Aug 77	Nov 78	15
Yemen	Aug 80	May 82	21
			Average elapsed time = 16 months

[a] Date recode tape with imputation available (from survey data sheets).

4 Measures and Strategies for the Assessment

The ideal measure of the benefit of editing, we suggest, would be the following ratio calculated for important diagnostic quantities such as proportions, means, correlation and regression coefficients, etc:

1—[(edited estimate – true value) / (unedited estimate – true value)], or equivalently, (unedited estimate – edited estimate) / (unedited estimate – true value).

Here the 'true value' refers to the sample rather than the population; editing should not be expected to compensate for sampling error. Unfortunately, of course, the true value is not known. The absolute value of the above ratio should lie in the range from zero to unity, with a value close to one indicating that editing is important, particularly if the denominator of the ratio is not small.

Lacking knowledge of the true value, two alternative measures could be considered. The first of these is the simple difference

(edited estimate – unedited estimate)

and the second is the difference relative to the presumably 'better' estimate, ie

(edited estimate – unedited estimate) / (edited estimate).

The relative difference will not in fact be employed, for several reasons. Most of the indicator quantities below will be proportions, and there is often ambiguity about whether the proportion or its complement is of interest. For example, if the edited data estimate that 20 per cent of all women are currently using contraception and 80 per cent are not, then unedited estimates of 21 and 79 per cent, respectively, will give a relative error of 1/20 or only 1/80, depending on an arbitrary choice as to which percentage is more interesting. Another consideration is that relative differences are most justified for ratio-level variables, ie for quantities with a natural zero. Many of the WFS quantities of interest do not have a natural zero, however, except perhaps in a very abstract sense. For example, the percentage of women who are currently pregnant in a WFS survey rarely falls outside the range of 10–14 per cent or so. Since a demographer could simply estimate the percentage pregnant to be within this range without ever conducting a survey, it is misleading to use a measure which implies that this quantity has a minimum possible value of zero. Hence the simple difference (edited estimate – unedited estimate) will be calculated and a judgement will be made as to whether this is large or small. Note that when the quantity is a percentage or proportion, it will be calculated on a base which excludes respondents for whom the relevant question was not asked (ie who have a 'not applicable' code).

When several categories of a variable are being compared, a simple summary measure of the net change in the distribution is the 'index of dissimilarity'. This is calculated as the sum of all positive changes, or, alternatively, as one-half the sum of the absolute values of all changes across categories, positive or negative. It can be interpreted as the minimum proportion of cases which would have to be shifted in one file in order to achieve the same distribution as the other file.

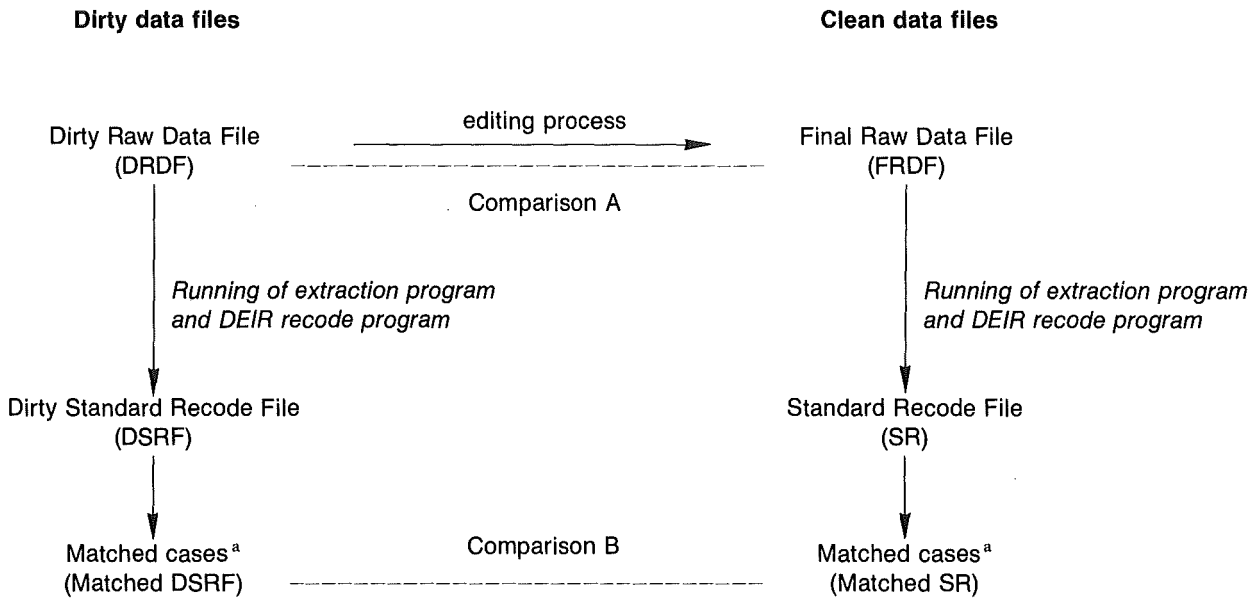
We now turn to the kinds of strategies which are feasible for this assessment. Some potential alternatives are closed to us because of their cost and because of the editing procedures actually adopted by WFS. For example, it is impossible to consider empirically any alternatives to the CONCOR computer program, and it is impossible to consider any procedures which require new references to the questionnaires. It is also impractical to develop alternative versions of the DEIR package.

The strategies which will be used are presented below in outline form, in terms of a raw data file which has not been cleaned, ie is 'dirty'.

A Compare the dirty raw data file (DRDF) and final raw data file (FRDF) on the relevant diagnostic indicators.

B If the DRDF were structurally correct, then one would apply the extraction program and DEIR to generate a dirty Standard Recode File (DSRF), and compare it with the clean one on the relevant diagnostic indicators. Since in general the DRDF includes structural errors, they must be compensated for in the following steps. First, prepare a new file which is a subset of the DRDF; refer to this as the matched DRDF. This will consist of those cases which (1) are complete in the DRDF—ie are structurally correct without any updating and (2) match in ID numbers with cases in the FRDF. Apply the extraction program to this file; call the result the matched dirty Standard Recode File or matched DSRF. Secondly, prepare a subset of the genuine Standard Recode File which consists of those cases which match the IDs on the matched DSRF. Call this the matched SRF. Then compare the matched DSRF and the matched SRF in terms of their values on the diagnostic indicators. Differences between these two files will be entirely due to the fact that the cases in the matched DSRF were not subjected to the range, skip, consistency, date, etc edit checks while in raw data form.

Part A of this outline indicates a comparison between the raw data files, whereas Part B refers to Standard Recode Files. Most of this report will deal with the latter. Figure 1 restates the logic behind the outline. In the diagram, the normal editing sequence would follow the arrows leading from the dirty raw data file to the right and then down to the Standard Recode File. In order to evaluate the editing process, we shall construct a so-called dirty



[a] Complete cases with IDs which match on both DSRF and SR.

Figure 1 Summary of editing strategy

Standard Recode File and compare it with the (clean) Standard Recode. Structural errors will be present in the dirty Standard Recode File, but are beyond our means to correct. Therefore, a structurally clean subset of the file will be constructed by matching cases with the clean file; this is an approximation to the file which would result if the structural editing were completed. Comparison between matched cases in the dirty and clean SR files demonstrates the omission of range, skip, and consistency editing.

Ideally, these comparisons would be based on a comparable sequence of tapes from a number of different countries representing a range of data quality. But because it is not standard WFS practice to save intermediate tapes during the editing (nor is it suggested that it should be, because of the typical in-country shortage of computer tapes), it will not be possible to be quite so systematic. Almost all early tapes presently available will be used, simply on a convenience basis.

5 Selection of Diagnostic Variables and Relationships

It is now necessary to identify a set of substantively significant proportions, etc which can be compared between the edited and unedited files. The number of variables and relationships which can be—and have been—extracted from a WFS survey is immense. That number must be pared down to those which in some sense have greatest interest. The country summaries and the First Country Reports will be used to make such a selection. Later, some important relationships at the second-stage analysis level will also be considered.

The First Country Report is required of each participating country and virtually always marks the first general release of survey results. When it is completed, there is typically a conference within the country to present the findings to a general audience; the country summary is prepared and distributed; plans are made for some second-stage analysis; and there is often a major transition in the staff who are assigned to the project on a day-to-day basis. The quite precisely specified tabulations of the First Country Report comprise the level in the analysis to which this assessment will give most attention.

The tabulation plan in the 'Guidelines for the First Country Report' is compactly restated in the 'Data Processing Guidelines' (Volume 2) in terms of the variables in the Standard Recode File. There are five major groups of tables, adding up to a total of 109 tables, each involving from two to five variables. (Knowledge and use of 15 specific contraceptive methods are recoded into binary variables. To avoid excessive weight on these two blocks of 15 variables, each block will be counted as one variable here.) Many tables are then repeated for each background variable in the standard set, but such tables are counted only once in the total of 109. Adding up the number of Standard Recode File references to specific variables (assuming that the fertility regulation module is being used, and not counting background variables), there are a total of 424 references to some 88 variables in this file. Many of these are closely related, eg are collapsed forms of other variables in the set or are simple constructions from two or more other variables. It is possible to identify a core of 13 out of the 88 which account for 322, or 76 per cent, of all 424 references to variables. These 13 are as follows, with the number of references to the variable or to other closely related variables given afterwards in parentheses:

(1) V011 Age in five-year groups (63)
(2) V108 Whether currently married (16)
(3) V110 Age at marriage in seven groups (22)
(4) V117 Marital duration in five-year groups (36)
(5) V206 Whether currently pregnant (24)
(6) V208 Number of children ever born (19)
(7) V213 Number of living children (43)
(8) V221 Number of living sons (11)
(9) V222 Number of living daughters (11)
(10) V402 Exposure status (32)
(11) V501 Desire for future birth (16)
(12) V511 Total number of children desired (20)
(13) V645 Pattern of contraceptive use (20)

Apart from these 13 variables and the ones which can be obtained directly from them, no other variables are referred to in more than six tables. Actually, most of the 88 variables are referred to only once or twice in the tabulation plan. We do not wish to suggest that the importance of a variable—even restricted to the level of the First Country Report—is proportional to the number of tables in which it appears. However, it could be argued that variables should be put in order of priority, at least in groups, with respect to the share of editing resources which they merit, and the variables which appear in more tables should have claim to more of these resources as well as being more indicative of the effects of editing.

Some background variables will be added to the above list. They receive little editing, in general, but because they are involved in many substantive conclusions it should at least be determined whether their distributions are affected. These variables differ somewhat from one country to another but always include the following:

(14) V702 Type of residence
(15) V704 Level of education
(16) V711 Last work status since marriage
(17) V804 Husband's occupation

Others are almost always used too—eg 'Region of residence', 'Religion', 'Ethnic group'—but they differ in importance and in number of categories from one country to another and will be excluded here.

For a final source of variables of major interest at the First Report level we have reviewed the country summaries to see whether some additional variables appear there. The following are found in several summaries:

(18) V223 Number of children born in the first five years of marriage
(19) V225 Number of children born in the past five years
(20) V231 Length of the closed interval (grouped)
(21) V233 Length of the open interval (grouped)
(22) V303 Duration of breastfeeding (grouped)
(23) V635 Current use of specific contraceptive methods
(24) V641 Use of any method in the closed interval
(25) V644 Use of any method in the open interval

Some variables are given in the Standard Recode File in both grouped and ungrouped form. For our purposes, the grouped form is somewhat easier to use. Age-specific

fertility rates are not included above as variables, but will be given below among the selected tables.

Thus we have a list of 25 variables of major interest. The marginal distributions of these may be compared at successive stages of editing to see net changes; one can also construct a cross-tabulation of an unedited versus edited version of the same variable to identify gross shifts in the distribution.

We now turn to the selection of diagnostic relationships among variables at the First Report level. Because of the number of relationships involved, it is necessary to reduce the list of standard tables in that report to a more manageable size. To begin with, we shall distil them to only two-way tabulations. The choice of these is based on a review of the country summaries, which present the findings of the First Country Report in vastly reduced form (eg approximately 12 pages of text). As in the case of the univariate selection, it must be emphasized that the selection is not claimed to identify the most important relationships as such. We are simply following a consensus among the writers of the country summaries, which it is hoped is shared by the users of these summaries, as to which relationships have the broadest interest. It is assumed that our conclusions for these tables can be generalized to others which might be more important for a particular researcher.

The following list includes tables most often used, as well as some which are only occasionally used. In addition, all the summaries give means, percentages, etc drawn directly from the list of principal variables given in the preceding section. The first four of the following two-way layouts refer to fertility; the next three to preferences; and the final three to contraceptive use.

(1)	V208 × V117	Children ever born by marital duration
(2)	V208 × V011	Children ever born by current age
(3)	V225 × V011	Children born in the past five years by current age
(4)		Age-specific fertility rates for past five years
(5)	V501 × V213	Desire for future birth by number of living children
(6)	V501 × V221 × V222	Proportion wanting another child by numbers of living sons and daughters
(7)	V511 × V011	Total desired family size by current age
(8)	V635 × V011	Current use of specific contraceptive methods by age
(9)	V231 × V641	Length of the closed interval by whether the couple used contraception in the closed interval
(10)	V645 × V011	Pattern of contraceptive use by current age

6 Case Studies of Raw Data Files

In order to convey some understanding of the volume and type of changes which are at issue, we shall begin the empirical investigation by looking in some detail at selected aspects of machine editing of the WFS surveys in Malaysia, Yemen, and Ghana.

6.1 CASE STUDY 1: MACHINE EDITING IN MALAYSIA

This section will be introduced with some data from the First Country Report on the 1974 Malaysian Fertility and Family Survey (MFFS). This was one of the very first surveys and did not use an editing package, and for that reason it does not typify the WFS experience. However, its First Country Report is the only one which provides any details on machine editing.

The MFFS included 6318 ever-married women. The editing was done in ten passes through the raw data; on the tenth pass no errors were detected. The errors or inconsistencies were only divided into the following simple classification: programming errors; edit specification errors; punching errors; and coding errors. The first two categories are by-products of the editing process itself, of course, and although inevitable, happened in Malaysia to produce more errors (spurious errors) than the other two sources. The percentage distribution across the four types was 61.4, 13.1, 6.9 and 18.6 per cent respectively, of a total of 25,012 errors. Ignoring the first two types, there were 6328 errors, of which 26.9 per cent arose in the key-punching and 73.1 per cent in the coding—that is, were present in the questionnaire themselves. This level of about one detected error per questionnaire is quite low by usual standards. A probable reason for the low level is that in Malaysia, all items were coded twice and then matched before date entry. The First Country Report also notes that the sixth to ninth passes through the data only produced a total of 19 additional errors. If the editing process had stopped after only three passes, it would have corrected 93.7 per cent of the 6328 errors, leaving only one error in any item for every 16 questionnaires. This limited but nevertheless unique breakdown provides an indication of the volume of editing which is at issue.

6.2 CASE STUDY 2: DATE EDITING IN YEMEN

Very limited analysis of the impact of editing in Yemen appears possible because of the absence of intermediate tapes. The first available tape in the sequence is IND2, which follows the consistency edit and is prior to the date edit. This will be compared with the final raw data tape, on which manual date edits have been made; the only differences will involve the date variables.

The net change between two marginal distributions on the two tapes is defined to be the minimum number of cases that would have to be shifted to other codes in order to achieve exact correspondence. If, say, one woman's age is changed from 29 to 30 in the editing, and another woman's age is changed from 30 to 29, then the two changes will cancel each other out and have no net impact on the marginal distribution. In analyses at the individual level, to be sure, there may well be associated changes which do not cancel out, because the respondents may differ in ways which are associated with their age (or whichever variable is under discussion). In a cross-tabulation, the probable effect of two changes such as the above would be to attenuate a relationship between age and another variable. However, net change in a marginal distribution is suggestive of the effect of editing upon more complex arrangements of the data.

Table 2 Changes resulting from manual date edit in Yemen

Question	Label change	Net change	Case base	D	Increase in sample size
Q107	Current age	20	2605	0.0077	0
Q108Y	Year of birth	20	2605	0.0077	1
Q203Y	Year started living with husband	22	2455	0.0090	4
Q2112Y	Year separated/husband died	17	526	0.0323	9
A, B, C, 326Y	Year of birth of child	81	9823	0.0082	4
	Child 1 – 4	43	6713	0.0064	1
	Child 5 – 9	29	2797	0.0104	3
	Child 10 +	9	313	0.0288	0
A, B, C, 331Y	Year of foetal losses	53	782	0.0678	9
	Losses before 3rd birth	25	397	0.0630	0
	Losses after 3rd birth	28	385	0.0727	9

For Yemen, the net changes from the manual date edit are so small that only a few variables justify any mention at all. These are shown in table 2. Note that D is the index of dissimilarity, the net change divided by the case base for the variable. The term 'increase in sample size' refers to the number of cases shifted out of a no-response category as a result of the manual edit. The table does not relate the changes to sampling error, does not give changes in parameter estimates such as the mean, and does not give the change in standard errors because all of these quantities would be so small.

The net change in the effective case base is below one per cent for all variables except (a) the year when separated or when the husband died—a variable which is only defined for about a quarter of all the women; (b) year of childbirth for later births—actually beginning with births of order 7; and (c) foetal losses. The higher level of shifting for foetal losses suggests that the structure of this part of the questionnaire was relatively deficient.

Not only were there very few changes in the detailed coding of these measures; the shifts were usually small and had only a minute impact on summary measures of the distributions. Thus, if the age categories are broadened to the familiar five-year groupings, then the number of net changes declines from 20 to only three cases out of 2605. It is pointless to compare means, variances, etc under such circumstances. Several other date-related variables also appeared in the Yemen survey, many from the FOTCAF module (on factors other than contraception affecting fertility). These variables were even less affected than those reported on above. The increase in the effective sample size was also small, never exceeding nine women.

According to data processing records, the date edit for Yemen proceeded in typical fashion. The error printout listed about 200 inconsistencies. Ten years of age was taken as the minimum possible age at marriage. Compared to other countries, this is a liberal minimum; if a higher age such as 12 had been selected, then more cases would have been listed.

The reported 200 inconsistencies were then reviewed in Yemen by a WFS Central Staff member and the local survey director, referring to the original questionnaires as necessary. Although it is not possible to be certain, a review of the error listing and resolutions (kept on file) suggests that approximately half of the inconsistencies were resolved without even going to the questionnaires. For example, if the birth years of successive children are given as '73, '47, and '75, then it is clear that a digit reversal occurred in the keypunching of the middle date, and it should be recoded to '74. With experience, one can correct the most common keypunch errors just from the error printout, without actually consulting the questionnaires.

The resolution of the 200 errors was completed in a week and two more weeks were required for the computer updating, which was done by batch mode in Yemen. This is standard for WFS, although in some settings at present, and presumably in many more in the future, interactive editing would save some time. As a conservative estimate, the manual date edit required three weeks of professional London staff time and six weeks of local professional time. This estimate does not include the preparation of the extraction program, because that would certainly have to be written under any plausible editing alternative so long as a Standard Recode File were to be created. Also, the extraction program was written in advance and did not itself cause any delay in the Standard Recode File.

In trying to determine whether any other parts of the process should have been bypassed because they were not cost-effective, one must compare what was done with plausible but cheaper alternatives. We have noted that the impact of editing was small—at least the impact on the marginal distributions—but we have also noted that the costs were small. Would the alternatives have reduced the personal time or elapsed time substantially?

In our view, the only step which it might have been reasonable to avoid was the examination of the questionnaires. About 100 inconsistencies were referred to the questionnaires; these were spread over many variables, including some of very little interest for the First Country Report. As it happened, it was not difficult to get access to the questionnaires in Yemen, and the trip there by WFS Central Staff was required for other purposes as well, but there is no evidence that this step in the editing was critical. Given the existence of the DEIR program, and the fact that perhaps half of the date inconsistencies could be resolved just from the error printouts, we infer that consultation of the questionnaires was the only step which it might have been acceptable to omit.

6.3 CASE STUDY 3: COMPARISON OF AN EARLY AND THE FINAL RAW DATA FILE FROM GHANA

Empirical evidence of net changes due to editing is available from the Ghana Fertility Survey (GFS) of 1978–79. We have access to both (a) an early file which appears to be quite dirty in terms of both structural and consistency problems, although we are not certain that it is the original raw data file, and (b) file IND3, the final raw data file, on which all updates through the date updates have been made. These will be referred to as RDBE and RDAE, for 'raw data "before" editing' and 'raw data "after" editing'. Comparisons will be made on the high priority variables for the Standard Recode File, or on similarly defined proxies, and on some of the most important background variables.

The first variable to be examined is age. Table 3 gives the marginal distribution of this variable in five-year groups, before and after editing.

Table 3 Ghana Fertility Survey, Q107: age distribution (in five-year groups) before and after editing

Age	RDBE	RDAE	Difference
15–19	1349	1351	+2
20–24	1202	1210	+8
25–29	993	993	0
30–34	821	821	0
35–39	691	691	0
40–44	579	588	+9
45–49	461	471	+10
Out of range	15	0	−15
Not stated	4	0	−4
Blank	16	0	−16
Total	6131	6125	−6

The table shows that six cases were dropped from the file, basically for structural reasons. The net increase in the case base was $6125-(6131-31-4)=29$ cases, a fraction $29/(6131-31-4)=29/6096=0.0048$, or about one-half of one per cent. The largest increase, both numerically and proportionately, was in the 45–49 age group—an increase of ten women. Is this increase noteworthy? The proportion of the sample who were aged 45–49 after the editing was 471 out of 6125, or 7.69 per cent. If this were a simple random sample, then the standard error of the percentage would be 0.34 per cent and a 95 per cent confidence interval for the percentage aged 45–49 in the population would range from 7.02–8.36 per cent. Actually, of course, the sample is not a simple random sample. The cluster design reduces the effective sample size, and the correct confidence interval has a slightly greater width. The percentage aged 45–49 that would have been calculated prior to editing, $100\times461/6096=7.56$ per cent, is only 0.13 per cent below the 'clean' estimate and is well within the range of the confidence interval, even if no allowance is made for the design effect.

We conclude that the editing was not particularly important for the marginal distribution of age. It increased the effective case base by only half a per cent, which would have reduced the standard error of estimates by only 0.24 per cent. In no cell was the change in the estimated proportion even close to the sampling error of the estimate.

The second variable to be considered from the Ghana Fertility Survey is the number of children ever born. Table 4 gives the complete distribution. The effective sample size increased from $6131-24=6107$ to 6125, ie an increase of 19 cases or 0.31 per cent. It may be noted that in the two-way relationship between completed family size and age, the loss of cases (ie the number of cases missing a valid code on one or both of the variables) would range between 25 and $25+19=44$. There are a number of shifts between the two distributions. The sum of the absolute deviations before and after editing is 46; there are 32 net increases and 14 net decreases. The only parities to have a net loss of cases were 0 and 10, 11, 12. Once again there is a curvilinear pattern to the adjustments, but it is the reverse of that found for age.

Table 4 Ghana Fertility Survey, Q213: number of children ever born, before and after editing

Number	RDBE	RDAE	Difference
Blank (none)	1529	1521	−8
1	893	900	+7
2	798	800	+2
3	665	672	+7
4	585	585	0
5	427	430	+3
6	369	371	+2
7	307	313	+6
8	227	228	+1
9	143	147	+4
10	104	102	−2
11	39	38	−1
12	10	7	−3
13	10	10	0
14	1	1	0
Out of range	24	0	−24
Total	6131	6125	−6

The change in the marginal distribution is small relative to the standard error. For example, the largest relative change is at parity 7. The standard error of the percentage at this parity is 0.28 per cent under the simple random sample model, and the actual standard error is considerably greater. The 'before' estimate of the percentage is 5.03, the 'after' estimate is 5.11 per cent. The difference, 0.08 per cent, is just a fraction of the standard error.

For a third illustrative variable from the Ghana Fertility Survey, we now consider 'total desired family size'. In Ghana this number tends to be rather large: half the women stated a desired family size of six or more children. The marginal distributions before and after editing are given in table 5. It was not clear what kinds of checks were used on this variable in order to generate error statements, apart from the review of cases which were coded 'out of range' or 'blank' (an illegal code). Almost all of the 15 out of range codes were obvious keypunch errors, viz codes 40, 50, 60, 70, and 80, which should have been 4, 5, 6, 7, and 8 respectively. The value of the highest allowable response was arbitrary; it was apparently set at 20, but we do not know if other high codes such as 15, 16, and 20 itself were checked. The 14 blanks were reassigned legal codes, but the basis for the reassignments is not known. In addition, there were 10 changes in the range of special codes 91–99 and 29 changes in the range 0–20. Of the latter group, 13 were changes out of code 0, which may have been specially checked against the questionnaire because it was felt that 0 was a very unlikely response in Ghana. As the table shows, after editing the childless category has become completely empty. It thus seems likely that some of the editing of this variable was done on an *ad hoc* basis—particularly the 29 changes to codes which were already in the legal range 0–20 and the reassignment of the 14 blank codes. This is disquieting because the reassigned codes from these sources tend to raise the mean of the distribution and tend also to reduce the dispersion.

We suspect a high level of response error for desired family size, best indicated by the marked heaping on 4, 6, 8, and 10 but also indicated by the high percentage who gave non-numeric responses. This sort of evidence implies low reliability and low face validity for the responses, exactly the situation in which the returns from resolution of a relatively small number of detectable errors can be reassuring but have no measurable benefits for the analysis. Even after editing, 628 women or 10.25 per cent of the sample still had special codes 91–99 and could not have been included in regressions, etc so the apparent gain of 34 cases in the rest of the sample, from $6115-652=5463$ to $6125-628=5497$, is negligible. We conclude that the value of this variable would not have been seriously impaired if the out of range and blank codes had been automatically collected under some default code.

We have reviewed a number of other key variables from the Ghana Fertility Survey with findings similar to those for age and children ever born or desired. In no case have we found a proportion which was substantially affected in terms of standard errors. For some variables, it is clear

Table 5 Ghana Fertility Survey, Q576: total number of children desired, before and after editing

Code	RDBE	RDAE	Difference	
0	13	0	−13	
1	10	10	0	
2	131	132	+1	
3	278	280	+2	
4	1669	1681	+12	
5	612	615	+3	
6	1320	1336	+16	
7	358	362	+4	
8	410	419	+9	
9	188	183	−5	
10	344	344	0	
11	51	49	−2	
12	66	66	0	
13	6	5	−1	
14	5	5	0	
15	6	6	0	
16	1	1	0	
20	3	3	0	
91	359	370	+11	Non-numerical response
92	222	225	+3	Non-numerical response
93	14	14	0	Non-numerical response
95	9	9	0	Non-numerical response
98	6	0	−6	Non-numerical response
99	13	10	−3	Not stated code
Out of range	15	0	−15	
Blank	14	0	−14	
Total	6115	6125	+10	

that a default was employed just as in the automatic alternative discussed earlier for the reconciliation of error printouts. For example, in the Ghana Fertility Survey, Q214 asked whether the woman was currently pregnant, with legal codes (1) Yes, (2) No, (3) Don't know. Virtually all of the 43 women who were out of range or blank in the 'before' file were placed into code 2, the 'No' category. Clearly, this placement was somewhat arbitrary but was probably motivated either by the belief that the question was skipped over by the interviewer when it was clear that the woman was not or could not be pregnant, in which case one kind of bias would have been reduced, or else by the belief that error would be minimized if the unknown cases were all allocated to the largest category. At any rate, this decision very slightly reduced the estimated proportion pregnant from 10.79 to 10.73 per cent. Since the standard error of the proportion, estimated as before, is 0.40 per cent, once again the editing produced a change which was only a fraction of the standard error.

Our conclusion from the univariate analysis of the Ghana Fertility Survey is that if (a) the 'out of range' codes for a variable had been transferred to a 'not stated' category and excluded from the case base for that variable, and if (b) the more complex consistency checks had not been made or had not been resolved in any way, then the marginal distributions would have been equally as valid as the edited distributions. There would, however, have been small inconsistencies between the case bases; for example, the number of women answering questions appropriate for pregnant women would not have been exactly the same as the number of women supposedly pregnant. A simple default could have reconciled inconsistencies of this type.

The preceding comparisons for Ghana were based on alternative versions of the raw data file. The great bulk of substantive research is actually based on the Standard Recode File, which has all of the woman's data on a single long record, together with a number of recodes and imputations as required in the birth and marriage histories. As stated earlier in this report, the impact of editing is best measured by comparing alternative versions of this file, an effort to which we now turn.

7 Univariate and Bivariate Comparisons for Six Standard Recode Files

In section 6 we indicated, primarily in the case study of Ghana, the levels and some possible sources of edit failures or detectable inconsistencies in the raw data files. We shall now compare dirty and clean (ie unedited and edited) versions of the Standard Recode File. Following the reasoning presented in section 4, we have constructed dirty Standard Recode Files (DSRFs) for six participating countries. In the present section they will be compared with the current Standard Recode Files for the respective countries in terms of the univariate distributions and bivariate relationships developed in section 6. These six DSRFs were prepared in London by the second author of this report. Because this comparison is being made with the current SR, which is not necessarily SR01, the evaluation will include the impact of any subsequent updates of SR01.

The six countries are Bangladesh, the Dominican Republic, Ghana, Haiti, Lesotho, and Portugal. They were not selected according to any particular scheme, but simply by the availability of an early unedited raw data file; these six comprise virtually all countries for which any such file could be identified. The actual dates or pedigrees of these files are not known, and it is possible that they have in fact been partially edited. With only one exception, however (Lesotho), they contain structural errors which are usually removed prior to the range, skip, and consistency editing. They all include codes out of range, which are the first kind of inconsistency to be removed in machine editing. They are thus early files indeed, even if not the very first ones to follow data entry.

As described in section 4, we have compensated for the incomplete structural editing simply by matching complete cases on the clean and dirty SR Files. This simple expedient costs relatively few cases for each country, as shown in table 6. Although the percentage of cases lost in the matching is negligible except for Ghana, where it is about six per cent, one should not conclude that the structural editing is unimportant; quite the contrary. As described early in this report, all of the structural editing was intended to precede the range, skip, and consistency editing.

Table 6 Complete cases matched on clean and dirty Standard Recode Files

Country	Number of cases in full clean SR	Number of cases in matched SR	% of cases lost by matching files
Bangladesh	6504	6353	2.32
Dom. Rep.	3115	3068	1.51
Ghana	6125	5746	6.19
Haiti	3350	3338	0.36
Lesotho	3603	3603	0.00
Portugal	5148	5119	0.56

The low percentage in the last column of the table indicates that most structural checks had indeed been made earlier in the lineage of these files and that very little structural error remains for potential discovery during the consistency editing stage.

The computer program used to create the matched 'dirty' Standard Recode Files (DSRFs) were already written and had been used in the countries on their cleaned raw data tapes. They were not modified to allow for inconsistencies on the dirty tapes. It would be possible to modify these programs so that, for example, all out-of-range codes would be collected into the 'not stated' code, and so on, but defaults such as this were not incorporated.

In the case study of the raw data file for Ghana, we indicated that one might regard sampling error as a rough guide to whether a discrepancy between the dirty and clean files is 'large' or not. If the difference is substantially greater than might have occurred between two independent cleaned samples, then it may be regarded as serious; conversely, if it is within the range of sampling error, then it is not serious. This report will not rely heavily on such a standard or guide. We shall simply indicate at this point the approximate magnitude of the standard errors for the percentages under investigation. In a sample of 6000 cases, a percentage of 50 per cent have a standard error of 0.6 per cent and a percentage of 5 per cent have a standard error of 0.3 per cent. In a sample of 3000 cases, the corresponding standard errors are 0.9 and 0.4 per cent. In applying this rule of thumb, one should of course consider that the effective sample sizes are a good deal smaller than stated because of the design effect, and that we are making multiple comparisons rather than individual ones.

7.1 CHANGES IN DISTRIBUTIONS

Table 7 indicates the differences in the percentage distributions between the matched dirty and clean SR Files for the six countries and for the 25 diagnostic variables. Two columns are given for each country. The first of these, headed 'Max', gives the absolute value of the maximum change in any category of the diagnostic variable. For example, in Bangladesh, the category of 'Age' (V011) which changed the most was age <20. In the matched clean SR, 22 per cent of the sample were given with age <20, and in the matched dirty SR, 22.5 per cent; the absolute net change was 0.5 per cent. This is the number given in the table. The second column in each pair, headed 'Mean', gives the average absolute change across all categories of the variable. In the case of Age in Bangladesh, to continue the example, there were changes of 0.1 per cent or more in three other age categories. Age 20–24 changed by 0.4 per cent; 30–34 by 0.1 per cent and 45–49

Table 7 Changes in distributions between the matched dirty and clean Standard Recode Files by country

Variable	Label	Country:	Bangladesh		Dom. Rep.		Ghana		Haiti		Lesotho		Portugal	
		Categories[a]	Max[b]	Mean[c]	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean
V011	Age (5 years)	7	0.5	0.2	0.1	0.0	0.2	0.1	0.1	0.0	0.3	0.1	0.1	0.1
V108	Currently married	2	no change		0.2	0.2	0.2	0.2	0.5	0.5	no change		0.3	0.3
V110	Age at marriage	5–6	0.2	0.1	0.3	0.1	1.7	0.7	0.9	0.4	0.8	0.4	0.3	0.2
V117	Yrs since marriage	6–7	0.2	0.1	0.3	0.2	0.6	0.3	0.3	0.2	0.2	0.1	0.1	0.0
V206	Currently pregnant	2	0.1	0.1	no change		0.1	0.1	0.1	0.1	no change		no change	
V208	No. children born	5–8	no change		no change		0.1	0.1	no change		0.1	0.0	no change	
V213	Living children	5–7	0.4	0.1	no change		0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0
V221	Living sons	4–6	0.3	0.1	no change		0.1	0.0	0.1	0.1	0.3	0.1	0.2	0.1
V222	Living daughters	4–5	0.4	0.2	no change		0.1	0.0	0.2	0.1	0.3	0.2	0.1	0.1
V223	Children in first 5 years	3–4	0.5	0.4	0.9	0.5	2.4	1.2	1.7	0.8	1.5	0.6	0.9	0.4
V225	Children in last 5 years	3–4	0.7	0.5	0.2	0.1	0.9	0.5	0.1	0.0	0.5	0.4	0.3	0.2
V231	LCBI	4–5	1.3	0.6	0.4	0.2	1.7	0.3	0.8	0.3	0.4	0.1	0.4	0.2
V233	Open interval	5	0.5	0.3	0.3	0.1	0.3	0.2	0.3	0.2	0.5	0.2	0.6	0.2
V303	Length breastfed	6–9	no change		no change		0.1	0.0	4.0	0.9	not used		no change	
V402	Exposure status	4–5	0.1	0.1	0.4	0.2	0.3	0.1	0.9	0.5	0.3	0.2	0.1	0.1
V501	Desire birth	3	no change		0.3	0.2	0.4	0.3	2.0	1.0	0.4	0.3	no change	
V511	Children desired	4–5	no change		no change		no change		no change		no change		no change	
V635	Method currently used	1–4	no change		0.4	0.2	0.3	0.3	1.2	0.8	0.7	0.7	0.2	0.1
V641	Type used in LCBI	2–3	0.5	0.4	0.6	0.4	0.3	0.2	not used		0.4	0.4	0.2	0.1
V644	Use FP in OBI	2–3	0.2	0.2	0.8	0.5	0.3	0.2	0.9	0.6	0.8	0.6	0.2	0.2
V645	Pattern of contraception	4–7	0.2	0.1	0.3	0.1	0.4	0.1	1.3	0.8	0.5	0.2	0.2	0.1
V702	Type of residence	1–3	no change		no change		0.2	0.1	no change		no change		not used	
V704	Level of education	3–6	0.1	0.0	no change		no change		0.4	0.2	no change		0.2	0.1
V711	Last work	2–4	0.1	0.0	0.2	0.1	0.4	0.2	0.5	0.3	0.1	0.0	0.2	0.1
V804	Partner's occupation	1–6	0.1	0.0	1.0	0.4	0.8	0.3	1.1	0.4	no change		no change	

[a] Omitting categories which include less than 5 per cent of the sample.
[b] Absolute value of the maximum change in any category.
[c] Average absolute change across all categories.

by 0.1 per cent. The average change across seven categories was therefore $(0.5+0.4+0.1+0.1)/7=0.157$, rounded to 0.2 per cent. The index of dissimilarity, if preferred, may be estimated as one-half of the mean deviation times the number of categories. Because of rounding in our figures, such a calculation will be approximate. Note that deviations below a rounded level of 0.1 per cent (0.05 per cent before rounding, or one case in 2000) are ignored, simply because they cannot be calculated with SPSS. Further, all calculations are rounded to the nearest 0.1 per cent. Categories including less than 5 per cent of the sample on both files are ignored.

For many variables it would be possible to calculate means and standard deviations, etc and differences between the clean and dirty versions in these summary statistics. We have not done this because changes would appear extremely small in that form. The selected indicators imply that the univariate distributions were affected very slightly, if at all, by the range, skip, and consistency editing. In our review of the 147 distributions (25 variables for each of six countries, except for three instances in which a variable was not used by a country), we found only 12 in which any category changed by more than one per cent as a result of editing. Three-quarters, or 112, of the distributions had no category which changed by as much as one-half of one per cent. Changes appear to be least likely for numbers of children ever born and living, fertility desires, and background variables. They appear to be more likely for variables which involve the birth history. There are exceptions to these generalizations, of course.

Table 7 shows that Portugal had no category which changed by as much as one per cent. The tapes from Bangladesh, Dominican Republic, and Lesotho had only one variable each in which any category experienced a net change of one per cent or more. Preliminary distributions from these tapes, particularly if rounded to the nearest one per cent, could have been issued with confidence. Ghana and Haiti had two and six variables, respectively, with deviations of one per cent or more. These are not trivial changes, and we shall comment briefly on each change which exceeded one per cent.

Changes in the Haiti SR file of one per cent or more

V223 (Children in first five years)—drop of 1.7 per cent in 'zero children' and increase of 1.2 per cent in 'two children'; mean increased by 0.036

V303 (Length of breastfeeding)—change in shortest durations, '0–2 months', from 6 to 2 per cent

V501 (Desire for future births)—'not stated' fell from 6.3 to 4.3 per cent, redistributed across More/No more/Undecided without changing their balance

V635 (Method currently used)—drop in 'none' of 1.1 per cent; increase in 'rhythm' of 1.2 per cent

V645 (Pattern of contraceptive use)—drop in 'used in closed interval' by 1.3 per cent

V804 (Partner's occupation)—increase of 1.1 per cent in the largest category, which is 'small farmer'.

Changes of one per cent or more in the remaining four countries

Ghana, V110 (Age at first marriage)—net shift of 1.7 per cent out of 'less than 15', redistributed across higher categories. If the youngest category is assigned to age 14.0, then the shift caused the mean to rise by 0.0875, or one month

Ghana, V223 (Children in first five years)—drop of 2.4 per cent in 'zero children', increase of 1.2 per cent in 'two children'; mean increased by 0.047

Bangladesh, V231 (length of LCBI)—small shift from '24–35 months' to '12–23 months'

Dominican Republic V804 (Partner's occupation)—shift from 'agricultural worker' to 'farmer'

Lesotho V223 (Children in first five years)—drop of 1.5 per cent in 'zero children', increase of 1.4 per cent in 'two children'; mean increased by 0.039

It will be noted from this list that V223 (Children in first five years) changed in exactly the same way in Ghana, Haiti, and Lesotho, with a small net transfer from zero to two and an increase in the mean of about 0.04 of a child. V804 (Partner's occupation) changed non-trivially in Dominican Republic and Haiti. No changes in background variables had been expected, but this one changed in a different way in these two cases. V635 (Method currently used) and V645 (Pattern of contraceptive use) both changed in Haiti. There was a tendency to shift away from never-use and toward current or recent use; we suspect that the changes were due to clarification of filters. The various other net transfers as a result of editing do not show a discernible pattern, and we have no reason to believe that they could have been anticipated.

7.2 CHANGES IN BIVARIATE ASSOCIATIONS AND FERTILITY RATES

Most net changes to the univariate distributions during the editing process were quite small. However, there will be many individual-level changes which cancel out in aggregations, not altering the distributions at all, and invisible to this point. The first point at which such changes will begin to appear is in cross-tabulations. We shall now turn to the ten two-way tabulations listed in section 5 of this report, looking for evidence of altered associations between variables as a result of the editing process. This review will include examination of age-specific fertility rates.

The association between the pairs of variables in the diagnostic tables will be summarized with the coefficient of association, C. This is a commonly used chi-square based measure; if X^2 is the calculated value of chi-square for a table of N cases, then

$$C = \sqrt{X^2/(X^2+N)}.$$

This measure is positive with limits of 0 and 1, and is designed to remove the effect of the sample size, N. Because it is based on a test statistic, X^2, and does not have a rationale in terms of proportionate reduction of error, for example, it is not an ideal measure of association. However, it has the advantage of being applicable to cross-clarifications of either nominal or ordinal variables, and it is not subject to sharp discontinuities for small interior changes in a table, which is a weakness of most alternative measures for cross-tabulations of nominal variables.

Table 8 Coefficients of contingency for two-way tables calculated from the matched dirty and clean Standard Recode Files for six countries

	Dirty	Clean	Diff	Dirty	Clean	Diff	Dirty	Clean	Diff
	Bangladesh			Dom. Rep.			Ghana		
V208 × V117	0.71241	0.71775	0.006	0.77221	0.77418	0.002	0.71924	0.72536	0.006
V208 × V011	0.70290	0.70505	0.002	0.67771	0.67851	0.001	0.71806	0.72316	0.005
V225 × V011	0.44462	0.44540	0.000	0.43126	0.43340	0.002	0.44277	0.45265	0.010
V501 × V213	0.49054	0.49704	0.006	0.49333	0.49284	0.000	0.45011	0.45275	0.003
V511 × V011	0.26468	0.26300	−0.002	0.42202	0.42477	0.003	0.44499	0.44544	0.000
V635 × V011	0.22304	0.22066	−0.002	0.35348	0.35132	−0.002	0.21818	0.21484	−0.003
V231 × V641	0.06569	0.05477	−0.011	0.14857	0.15696	0.008	0.32016	0.29719	−0.023
V645 × V011	0.37017	0.36974	0.000	0.40432	0.40429	0.000	0.35387	0.36303	0.009
	Haiti			Lesotho			Portugal		
V208 × V117	0.77427	0.77396	0.000	0.69579	0.69867	0.003	0.53795	0.53916	0.001
V208 × V011	0.66840	0.66784	−0.001	0.66087	0.66351	0.003	0.45831	0.46210	0.004
V225 × V011	0.43405	0.43955	0.006	0.40904	0.41149	0.002	0.46835	0.47093	0.003
V501 × V213	0.52417	0.52271	−0.001	0.40332	0.41232	0.009	0.48872	0.48878	0.000
V511 × V011	0.34517	0.34363	−0.002	0.31364	0.31451	0.001	0.25340	0.25192	−0.002
V635 × V011	0.25642	0.26902	0.013	0.14511	0.21287	0.018	0.31580	0.31801	0.002
V231 × V641	N.A.			0.07664	0.08004	0.003	0.27057	0.27503	0.004
V645 × V011	0.27888	0.26259	−0.016	0.44993	0.45113	0.001	0.26414	0.26597	0.002

Table 8 gives the values of C for the dirty and clean files, and the differences, for each of the six countries. These are computed for the ten two-way tables listed in section 5 except for the age-specific fertility rates (to be discussed below) and the proportion wanting another child by numbers of living sons and daughters. This latter table of proportions is derived from a three-way table of frequencies; we calculated C in each panel and reached the conclusion that the difference between the dirty and clean versions of this table was in the range of differences for other tables within each country. Those more detailed calculations are therefore omitted from table 8.

There are 47 pairs of coefficients in table 8, taking a wide range of values from nearly 0.0 to nearly 0.8. The differences within a pair, ie the clean C minus the dirty C, so to speak, are generally quite small. Fully 41 or 87 per cent of the pairs have a difference of less than 0.010 between the two values of C. The six largest differences (in absolute value) are 0.010, 0.011, 0.013, 0.016, 0.018 and 0.023. We do not believe that any researcher would have altered his or her inferences about these associations as a result of the data editing. For example, the largest difference in this set, 0.023, occurred for V231 × V641 in Ghana. In the dirty table, C was 0.320, and after editing it fell to 0.297, a relative decline of only seven per cent.

Earlier we mentioned the possibility that associations might tend to be strengthened by the editing, ie they might be somewhat attenuated unless this step was carried out. There is evidence here to support such an hypothesis. Out of the 47 differences given in table 8, seven are zero (to three decimal places), 11 are negative, and 30 are positive. This tendency for C to increase is statistically highly significant with a sign test. Further examination shows, however, that most of this imbalance can be traced to Lesotho. In this data set, C increased after editing in all of the diagnostic tables, but by an amount ranging from 0.001 to 0.018 and averaging only 0.005. Considering that the edited values of C ranged from 0.007 to 0.696 for that country, an average attenuation of 0.005 has no importance for the analysis.

To summarize, the small disturbances in univariate distributions are no more serious when we take a pair of variables together—at least not to the point of affecting interpretations. There is some evidence that the dirty coefficients tend to be slightly smaller than the clean ones, but by a trivial amount.

Fertility rates for the previous five years were calculated for each of the six countries using FERTRATE, the standard WFS computer program for this purpose. These were calculated for five-year age groups 15–19, ..., 45–49 and for the total (multiplied by five, the total fertility rate). We examined the rates for the clean and dirty SR Files and also the numerators and denominators of those rates, which measure numbers of births and woman-years of exposure. Table 9 gives the ratio of the unedited to edited estimates of these rates for the five years before the survey.

With the exception of Ghana, current fertility rates calculated from the dirty matched Standard Recode File are quite close to those from the clean matched file. The total fertility rate for years 0–4 before the survey is low by only one per cent in the dirty file for the Dominican Republic; low by two per cent for Bangladesh, Lesotho, and Portugal, and low by five per cent for Ghana. The five-year age-specific fertility rates are also generally low, and only occasionally too high. This pattern carries over to the numerators and denominators of the fertility rates; both are usually lower than in the edited data, but the loss to the numerator is a bit greater, pushing the rates down. There is no clear indication of any pattern across ages.

The dirty TFRs are generally a quite good approximation to the clean ones, in our view, but one could hope that the correspondence in age-specific rates was less erratic. The general deficiency in the numerators and denominators of the rates appears not actually to arise within the data,

Table 9 Ratio of unedited to edited estimates of age-specific rates and total fertility rates for years 0–4 before the survey for six countries

Age (yrs)	Bangladesh	Dom. Rep.	Ghana	Haiti	Lesotho	Portugal
15 – 19	1.00	0.99	1.00	1.00	0.99	0.99
20 – 24	0.98	1.00	0.97	0.99	1.00	0.98
25 – 29	0.97	0.99	0.94	0.97	0.97	0.98
30 – 34	0.99	0.98	0.94	0.97	0.99	0.98
35 – 39	0.96	0.97	0.94	0.94	0.94	0.97
40 – 44	0.91	1.00	0.91	0.99	0.96	0.90
45 – 49	1.20	1.00	0.96	1.00	1.07	1.10
TFR	0.98	0.99	0.95	1.00	0.98	0.98

however, but in how the data are processed by FERTRATE. The computer program assumes that all data presented to it have been completely cleaned and contain no '99' codes, etc. It includes no internal checks of its own and therefore has no default procedures. If the program encounters certain impossible values for the woman's birth date (or marriage date, in the case of marital fertility rates), such as a birth date which would give her an impossibly young age, then the woman will be dropped altogether and all of her births will also be dropped. Such a case will lead to a deficit in both the numerators and denominators of the age-specific rates. Then again, if an error is encountered in the birth history, such as a childbirth which was impossibly long ago, then the birth will be dropped, leading to a deficit in the numerator only. Like all fertility rate programs, FERTRATE works by accumulating an array of births and an array of exposure, and dates which fall outside of the dimensions of the array will be omitted entirely.

In our view, this handicap in FERTRATE (a handicap only if the data have not been completely cleaned) could be largely overcome by introducing defaults. At the very least, a count of unacceptable codes could be kept and then be used to inflate all rates. In other words, software modification could at least partially substitute for complete date editing. Yet, as was described earlier in this report, date editing with DEIR can be carried out quite expeditiously (recall that in Yemen it required only nine person-weeks). We therefore suggest that when DEIR or the equivalent is available, then the dates should be edited rather early. Apparently an aggregate measure of current fertility such as the total fertility rate can usually be computed safely before editing.

8 The Effect of Editing upon Multivariate Analyses

We now turn to our most stringent evaluations of the importance of range, skip, and consistency editing. Two multivariate analyses will be carried out identically on the clean and dirty versions of the Standard Recode Files for each of the six countries under study. Because each analysis involves several variables and rather complex calculations, there are many ways in which shifts which had a minor impact on univariate distributions and two-way cross-tabulations may now be quite important.

The two analyses range in complexity between the sophisticated use of a many-way table of the sort found in the standard tabulation plan, and a micro-level multiple regression typical of second stage analysis. We shall begin with the somewhat simpler example.

8.1 AN ANALYSIS OF CONTRACEPTIVE USE

The first analysis is simpler in the sense that it is based upon a many-way tabulation. It will be subjected to logit regression (using the GLIM computer program), but it involves no interval-level variables, in contrast to the second example. Our interest is in predicting current use of contraception among exposed women, using desire for another child, number of living children, and a background variable. The dirty and clean tapes will be compared in terms of whether they both lead to the same model (ie the same interaction terms) and the degree of correspondence between estimated coefficients for a particular model. The analysis is confined to currently exposed women (V404 = 1) partly because these were the only women who were actually asked about current contraceptive use and partly because this control imposes another demand upon the correspondence between clean and dirty files. (Note that, as is standard practice for WFS, sterilized women are counted as current users and as currently exposed.) The variables involved are:

V637 Current use of contraception (Yes/No)
V501 Desire for future birth (Wants more/Wants no more/Undecided)
V217 Number of living children (0–3/4–6/7 or more)
V711 Woman's work status since marriage (collapsed to No Work/Family–Self/Others)

The table, V637 × V501 × V217 × V711 for V404 = 1, was run on SPSS. It is similar to some tables in the First Country Report's standard tabulation plan, but is not precisely identical to any of them. The table was submitted to GLIM for estimation of all hierarchical models on the log odds of Yes vs. No on V637. This procedure was applied to the dirty and clean matched Standard Recode Files for each of the six available countries.

The dirty and clean files will first be compared in terms of which model best fits the dependent log odds on V637. These models will be stated in the notation of GLIM, with V501, V217, and V711 referred to by the symbols 1, 2, and 3 respectively. Then, for example, the model 1 + 2 + 3 + 12 denotes the fitting of the log odds on V637 using main effects for V501, V217, and V711 plus the two-way interaction of V501 with V217. Table 10 gives the optimal models for each file, ie out of all hierarchical models of main effects plus interactions, the ones which fit acceptably in terms of chi-square and use the smallest number of parameters.

The two files lead to exactly the same model in three cases: Bangladesh, Portugal and Ghana. In the other three countries (the Dominican Republic, Haiti, and Lesotho) the only difference is in the addition or deletion of a single type of interaction term, either 12 or 23. For all these cases, the optimal model in the clean file was acceptable but not optimal in the dirty file. That is, if the analysis had been limited to the dirty files, then in all six countries the researchers would have accepted the models which appear best in the clean file, but in only three of the six countries would the same model have been optimal.

Generally, one is less interested in the best-fitting model, as such, and more interested in the magnitude and significance of specific effects. Table 11 gives the estimated effects and their standard errors from the optimal models listed above. The table does not label the effects, because that is not important for the present purpose, but the first effect is the overall main effect; the next three pairs refer to variables 1, 2, and 3 respectively, and the subsequent groups of four refer to interactions 12, 13 and 23, respectively. Effects are calculated by GLIM as deviations from a reference category, which is always taken to be category 1. Thus, these are logit regression coefficients with dummy variable coding, rather than MCA-type coefficients which would add to zero in groups.

An asterisk is placed next to each coefficient which is greater (in absolute value) than 1.96 times its standard error. Such effects are nominally significant at the 0.05 level, but the true level of significance is somewhat less

Table 10 Optimal hierarchical models on the log odds of Yes vs. No on V637 for six countries

Country	Dirty file	Clean file
Bangladesh	1 + 2	1 + 2
Dom. Rep.	1 + 2 + 3	1 + 2 + 3 + 12
Ghana	1 + 2 + 3 + 12 + 13	1 + 2 + 3 + 12 + 13
Haiti	1 + 2 + 3 + 12	1 + 2 + 3
Lesotho	1 + 2 + 3 + 13	1 + 2 + 3 + 13 + 23
Portugal	1 + 2 + 3 + 13 + 23	1 + 2 + 3 + 13 + 23

Table 11 Comparison of the best-fitting logit regression models for six countries, as computed by GLIM from the dirty and clean matched Standard Recode Files

	Dirty (1+2)		Clean (1+2)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
A Bangladesh						
1	−3.804	0.2393	−3.857*	0.2461	−0.0530	−0.2215
2	1.841*	0.2509	1.862*	0.2576	0.0210	0.0837
3	0.3886	0.4333	0.3943	0.4366	0.0057	0.0132
4	0.1853	0.1220	0.2339	0.1226	0.0486	0.3984
5	0.3885*	0.1609	0.3887*	0.1621	0.0002	0.0012

	Dirty (1+2+3)		Clean (1+2+3+12)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
B Dom. Rep.						
1	−1.473*	0.1554	−1.518*	0.1702	−0.0450	−0.2896
2	1.936*	0.1713	1.872*	0.2173	−0.0640	−0.3736
3	0.3268	0.3482	0.8510	0.5084	0.5242	1.5055
4	−0.2062	0.1659	−0.2052	0.3315	0.0010	0.0060
5	−0.6197*	0.1912	−0.3845	0.5086	0.2352	1.2301
6	0.1664	0.2063	0.2621	0.2069	0.0957	0.4639
7	0.4226*	0.1511	0.4285*	0.1509	0.0059	0.0390
8	—	—	0.1463	0.3851	—	—
9	—	—	−0.1482	0.5520	—	—
10	—	—	−0.2303*	0.7457	—	—
11	—	—	−2.026	1.245	—	—

	Dirty (1+2+3+12+13)		Clean (1+2+3+12+13)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
C Ghana						
1	−3.146*	0.4583	−2.924*	0.4209	0.2220	0.4844
2	2.101*	0.8453	1.754*	0.8230	−0.2560	−0.3029
3	0.1223	0.9893	−0.3773	0.9622	−0.4996	−0.5050
4	−0.1830	0.1698	−0.1684	0.1676	0.0146	0.0860
5	0.1874*	0.3391	0.1793	0.3387	−0.0081	−0.0239
6	1.077*	0.4643	0.8424*	0.4274	−0.2346	−0.5053
7	1.968*	0.4935	1.797*	0.4574	−0.1710	−0.3465
8	0.01885	0.4185	0.05606	0.4071	0.0372	0.0889
9	−0.4544	0.5356	−0.4141	0.5263	0.0403	0.0752
10	1.551*	0.6403	1.679*	0.6325	0.1280	0.1999
11	−0.8874	1.205	−0.8455	1.203	0.0419	0.0348
12	−1.231	0.8145	−0.9905	0.7933	0.2405	0.2953
13	−0.6889	0.8821	−0.4756	0.8578	0.2133	0.2418
14	−1.717	0.9369	−1.283	0.9112	0.4340	0.4632
15	−0.2307	1.079	0.1431	1.061	0.3738	0.3464

Table 11 *(cont)*

	Dirty (1+2+3+12)		Clean (1+2+3)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
D Haiti						
1	−1.331*	0.1949	−1.335*	0.1885	−0.0040	−0.0205
2	0.8535*	0.2003	0.9224*	0.1772	0.0689	0.3440
3	0.2324	0.3268	−0.1327	0.2879	−0.3651	−1.1172
4	−0.09345	0.4703	−0.1836	0.1707	−0.0902	−0.1917
5	−8.777	40.01	0.3905	0.2317	9.1675	0.2291
6	−0.3407	0.1818	−0.2984	0.1803	0.0423	0.2327
7	0.3143	0.2259	0.3451	0.2231	0.0308	0.1363
8	−0.02894	0.5093	—	—	—	—
9	9.343	40.01	—	—	—	—
10	−1.605	0.9123	—	—	—	—
11	−0.06602	58.43	—	—	—	—

	Dirty (1+2+3+13)		Clean (1+2+3+13+23)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
E Lesotho						
1	−3.225*	0.1717	−2.987*	0.1594	0.2380	1.3862
2	1.500*	0.2569	1.592*	0.2410	0.0920	0.3581
3	−6.918	34.03	2.309*	0.7251	9.2270	0.2711
4	0.2440	0.2209	−0.003905	0.2376	−0.2479	−1.222
5	0.2236	0.3639	−0.1184	0.4363	−0.3420	−0.9398
6	1.027*	0.3238	2.845*	0.9587	1.8180	5.6146
7	0.3384	0.3970	−0.3336	0.4953	−0.6720	−1.6927
8	−1.448*	0.7210	−2.849*	0.9811	−1.4010	−1.9431
9	0.3653	0.5157	0.1575	0.5391	−0.2078	−0.4029
10	9.737	34.05	−11.61	72.47	−21.347	−0.6269
11	0.2392	80.05	−8.554	72.47	−8.7932	−0.1098
12	—	—	−1.403	1.001	—	—
13	—	—	1.096	0.5847	—	—
14	—	—	No cases		—	—
15	—	—	1.238	0.8969	—	—

	Dirty (1+2+3+12+13)		Clean (1+2+3+12+13)		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
F Portugal						
1	0.8057*	0.1216	0.8014*	0.1211	−0.0043	−0.0354
2	0.5756*	0.1545	0.5838*	0.1542	0.0082	0.0531
3	−0.3677	0.3474	−0.3505	0.3477	0.0172	−0.3505
4	0.2755	0.2618	0.2620	0.2621	−0.0135	−0.0516
5	−0.5913	0.4052	−0.7486	0.3973	−0.1573	−0.3882
6	−0.9665*	0.1754	−0.9736*	0.1740	−0.0071	−0.0405
7	0.2625	0.1555	0.2859	0.1552	0.0234	0.1505
8	0.6303*	0.2188	0.6324*	0.2177	0.0021	0.0096
9	0.4225*	0.2021	0.3960*	0.2018	−0.0265	−0.1311
10	0.1262	0.4745	0.05356	0.4723	−0.07264	−0.1531
11	0.3239	0.4877	0.3131	0.4868	−0.0108	−0.0221
12	1.002*	0.3171	−1.050*	0.3165	−0.048	−0.1512
13	−0.6195	0.3441	−0.6993*	0.3409	−0.0798	−0.2319
14	−0.5656	0.5118	−0.3370	0.5086	0.2286	0.4467
15	−0.2253	0.5870	0.1087	0.5951	0.334	0.5690

because of the design effect. We assume that the basic conclusions of the comparison are not affected by the use of a nominal level of significance.

The overlap between the dirty and clean files in their inferences about the six countries in question is indicated below.

		Clean file: Sig	Clean file: Not sig
Dirty file	Sig	23	1
	Not sig	3	87

Excluding the saturated model, which would fit each file perfectly, there are 19 possible effects for each country, or a total of $6 \times 19 = 114$. Generally there is good agreement; in only four out of the 114 possible inferences is there disagreement. Two of these were for the Dominican Republic, one was for Lesotho and the other was for Portugal.

One may ask whether there is any bias or attenuation in the magnitude of the coefficients. There are 60 pairs of estimates which can be compared (in the other 54, one or both of the estimates was so close to zero that it was excluded from the best fitting model). In 32 of these pairs, the clean file produces coefficients which are closer to 0, and in the remaining 28 the clean estimate is further from zero. If the dirty estimates were attenuated, then the balance would be in the opposite direction. In any event, with 60 pairs a 32:28 split is not significantly different from an even split. Similarly, there is no indication of a bias. Of the 60 pairs, the clean estimate is smaller than the dirty one 27 times and larger 33 times. Among the coefficients which are significant in a pair of files, the corresponding split is 12:11, as close to equality as possible.

The final comparison between the dirty and clean logit regressions will be based on the last column of table 11. The indices in that column are the difference between estimates (the clean estimate minus the dirty estimate), divided by the standard error of the dirty estimate. These indices, or standardized differences, as they will be called, are motivated by the following reasoning. Suppose that the data had not been cleaned, and the estimates from the dirty file had been accepted for analysis. Then the point estimate given in the first column of table 11, together with the standard error of that estimate, given in the second column, could be used to construct an interval estimate. A standardized difference is the necessary width of such an interval, expressed in standard deviations, if it is to encompass the clean estimate. It is not to be regarded as a test statistic in any sense, but rather as a measure of the accuracy of the dirty coefficients as estimates of the clean coefficients. Alternative indices could be devised, using either the dirty or clean standard error or both, with similar empirical results.

Out of the 60 standardized coefficients which it is possible to compute, 47, or 78 per cent, are in the range of -0.5 to $+0.5$. In these 47 cases, the clean coefficient is less than half a standard deviation away from the dirty estimate. In another five comparisons, the absolute value of the index is between 0.5 and 1.0 standard deviations away; in four more, between 1.0 and 1.5; in three more between 1.5 and 2.0; and in one case it is more than 2.0 standard deviations away. The most extreme case is for coefficient no 6 in the Lesotho data; and despite the differences in magnitude of the two estimates for this coefficient, both concur in attaching a high level of significance to this effect. The second and third largest standardized indices are also for Lesotho (coefficients no 8 and no 7 respectively) and they also do not entail any disagreement about significance.

Our interpretation of table 11 is that an analyst who was sensitive to the importance of sampling error would not have reached any notably different conclusions from the dirty files for these countries than from the clean files. This conclusion is based on examination of model selection, possible attenuation and bias, and the magnitude of deviations between the two estimates of each kind of effect.

8.2 AN ANALYSIS OF CURRENT FERTILITY

Our second example places substantially heavier demands upon the data. It requires processing of the complete individual-level files because it includes three interval-level variables as well as two categorical socio-economic variables and uses multiple regression. Moreover, all three of the interval-level variables are drawn from the dates in the birth and marriage histories. We are not aware of any published analysis of WFS data which uses exactly the same model, but there is close correspondence to a model used by Rodríguez and Cleland (1981).

As this analysis includes the effect of date imputation it is necessary to discuss how far this imputation will be the same in the dirty and clean tapes. Imputation is based on pseudo-random numbers generated from a seed, which is always the same seed. Therefore the sequence of numbers is always the same. However, they will not produce the same individual date imputations for each event in the two files because once one of the files has an event to be imputed which the other file does not have they will get out of step. This, in general, will happen quite early on. Thus the two sets of imputations may be regarded as almost independent, though based on the same imputation principles.

Our interest here, as in the preceding example, is almost entirely in whether the clean and dirty files yield comparable conclusions about the adequacy of a model, and not in the model itself. Nevertheless, we have tried to select a model which is plausible and of some general interest. We shall take the number of births in the preceding five years as the dependent variable, and regress it upon marital duration, age at marriage, two background variables (represented as dummy variables) and interactions between duration and the background variables. Women married less than five years will be excluded. In particular, we use:

Y = V225	Births in the past five years
D = V116	Years since first marriage
A = V109	Age at first marriage
U (from V702)	Dummy variable for urban women
E1 (from V704)	Dummy variable for lower primary education
E2 (from V704)	Dummy variable for upper primary education
E3 (from V704)	Dummy variable for secondary education and above.

The omitted or reference category for 'type of place of residence' is 'rural' and for 'education' is 'none'. The regression equation is

$$Y = b_0 + b_1(D) + b_2(A) + b_3(U) + b_4(DU) + b_5(E1) + b_6(DE1) + b_7(E2) + b_8(DE2) + b_9(E3) + b_{10}(DE3)$$

This regression was prepared with SPSS from the dirty and clean files for the same six countries as above. The regression coefficients, standard errors, and R^2 values from these runs are presented in table 12. Because the calculation of standard errors by SPSS does not take account of the design effect, as mentioned earlier, all of the estimates of standard errors are conservative. Nevertheless, an asterisk is again attached to each coefficient which is at least 1.96 times its estimated standard error in order to suggest which effects are statistically significant. The phrase 'Not in the equation' appears where the coefficients were automatically omitted by SPSS because they were very small relative to their standard errors.

The entries in table 12 will be examined systematically in very much the same way as the logit regression estimates in table 11. We shall look at the overlap in significance of dirty and clean estimates; possible attenuation in the dirty estimates; possible bias; and the standardized differences between the two sets of differences.

Out of the 60 regression coefficients for the six countries (ignoring the constant term, b_0, which is always significant), 18 are significant in both files at the nominal 0.05 level and another six are significant in one file but not the other. There are three discrepancies for Portugal, involving b_5, b_6 and b_{10}; two discrepancies for Bangladesh, involving b_3 and b_7, and one for Lesotho involving b_3 again.

		Clean file	
		Sig	Not sig
Dirty file	Sig	18	2
	Not sig	4	36

b_3, for Bangladesh, involves the largest among all 60 differences. The two estimates are of the same sign, but differ greatly in magnitude. The other discrepancies in putative significance are due to sharply different estimated standard errors. The coefficients themselves are not greatly different. The remaining 36 coefficients are insignificant in both files.

Next we consider the possibility that the dirty coefficients are attenuated, ie are systematically closer to zero than the clean estimates. Out of the 49 pairs of coefficients in which SPSS produced both a dirty and a clean estimate, the dirty estimate was closer to zero 26 times. The balance is far from significant with a two-tailed sign test.

The difference in column 5 of the table, the clean estimate minus the dirty estimate, is positive for 20 and negative for 29 pairs. That is, more often than not, the clean estimate is more positive than the dirty. There is perhaps a suggestion of bias here, but it is again far from significant with a sign test.

Finally, the standardized differences in column 6 of table 12 will be reviewed. To repeat the earlier rationale for dividing the differences in column 5 by the estimated standard errors of the dirty coefficients, these ratios give the width of a dirty interval estimate, in terms of standard deviations, which would be required to enclose the clean estimate. Expressed in this way, the discrepancies are similar to those found in the logit regressions. Fully 38 out of 49, or 78 per cent of the standardized differences, are less than 0.5 in absolute value; six are between 0.5 and 1.0 standard deviations away; three are between 1.0 and 1.5; and two are over 2.0. The two largest pertain to b_1 and b_{10} for Portugal. The dirty and clean estimates for these coefficients do not differ greatly but the standard errors of the dirty estimates are relatively small and produce large standardized differences.

To summarize the regression results in table 12, most countries are virtually identical in the dirty and clean versions. An analyst who was sensitive to the effect of sampling error would have reached conclusions from the dirty file which would be altered only trivially, if at all, by the cleaning.

Table 12 Comparison of the regressions for six countries, as computed by SPSS from the dirty and clean matched Standard Recode Files

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
A Bangladesh						
b_1	−0.0316*	0.0093	−0.0441*	0.0015	−0.0125	−1.3441
b_2	−0.0154*	0.0061	−0.0202*	0.0061	−0.0048	−0.7869
b_3	0.3828*	0.1723	0.1417	0.1025	−0.2411	−1.399
b_4	−0.0127	0.0094	−0.0079	0.0055	0.0048	0.5106
b_5	0.0589	0.0477	0.0578	0.0463	−0.0011	−0.0231
b_6	Not in the equation		Not in the equation		—	—
b_7	0.1616	0.0881	0.1143*	0.0420	−0.0473	−0.5369
b_8	−0.0027	0.0051	Not in the equation		—	—
b_9	Not in the equation		0.1373	0.1742	—	—
b_{10}	Not in the equation		−0.0088	0.0119	—	—
R^2	0.1783		0.1857		0.0074	

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
B Dom. Rep.						
b_1	−0.0816*	0.0073	−0.0810*	0.0072	0.0006	0.0822
b_2	−0.0379*	0.0067	−0.0393*	0.0067	−0.0014	−0.2090
b_3	−0.5418*	0.1128	−0.5300*	0.1132	0.0118	0.1046
b_4	0.0052	0.0063	0.0055	0.0064	0.0003	0.0476
b_5	0.0166	0.1719	0.0515	0.1705	0.0349	0.2030
b_6	0.0042	0.0084	0.0025	0.0084	−0.0017	−0.2024
b_7	−0.3454	0.1766	−0.3294	0.1766	0.0160	0.0906
b_8	0.0157	0.0095	0.0142	0.0095	−0.0015	−0.1579
b_9	−0.5971*	0.2056	−0.5362*	0.2070	0.0609	0.2962
b_{10}	0.0151	0.0116	0.0116	0.0118	−0.0035	−0.3017
R^2	0.2822		0.2819		−0.0003	

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
C Ghana						
b_1	−0.0471*	0.0024	−0.0476*	0.0024	−0.0005	−0.2083
b_2	−0.0085*	0.0043	−0.0104*	0.0043	−0.0019	−0.4419
b_3	−0.1724*	0.0722	−0.1499*	0.0718	0.0225	0.3116
b_4	0.0055	0.0044	0.0040	0.0043	−0.0015	−0.3409
b_5	−0.1522	0.1831	−0.1248	0.0819	0.0274	0.1496
b_6	0.0015	0.0117	Not in the equation		—	—
b_7	Not in the equation		0.0313	0.1252	—	—
b_8	−0.0019	0.0037	−0.0027	0.0082	−0.0008	−0.2162
b_9	−0.0449	0.0779	−0.0774	0.0789	−0.0325	−0.4172
b_{10}	−0.0102	0.0057	−0.0060	0.0057	0.0042	0.7368
R^2	0.1395		0.1414		0.0019	

Table 12 *(cont)*

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
D Haiti						
b_1	−0.0558*	0.0038	−0.0571*	0.0038	−0.0013	−0.3421
b_2	−0.0249*	0.0053	−0.0248*	0.0053	0.0001	0.0189
b_3	−0.4370*	0.1198	−0.4692*	0.1199	−0.0322	−0.2688
b_4	0.0060	0.0071	0.0069	0.0071	0.0009	0.1268
b_5	−0.1196	0.1429	−0.1268	0.1456	−0.0009	−0.0693
b_6	0.0051	0.0092	0.0036	0.0093	−0.0015	−0.1630
b_7	−0.2373	0.1881	−0.1985	0.1873	0.0388	0.2063
b_8	0.0063	0.0111	0.0040	0.0111	−0.0023	−0.2072
b_9	−0.0242	0.2242	−0.0987	0.2251	−0.0745	−0.3323
b_{10}	−0.0168	0.0156	−0.0078	0.0161	0.0090	0.5769
R^2	0.1762		0.1805		0.0043	

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
E Lesotho						
b_1	−0.0530*	0.0071	−0.0553*	0.0071	−0.0023	−0.3239
b_2	−0.0245*	0.0054	−0.0222*	0.0053	0.0023	0.4259
b_3	−0.1576*	0.0644	0.1642	0.1410	−0.0066	−0.1025
b_4	Not in the equation		0.0011	0.0080	—	—
b_5	−0.0438	0.1776	−0.1114	0.1740	−0.0676	−0.3806
b_6	0.0033	0.0089	0.0075	0.0088	0.0042	0.4719
b_7	0.1390	0.1495	0.1138	0.1457	−0.0252	−0.1686
b_8	−0.0036	0.0077	−0.0016	0.0076	0.0020	0.2597
b_9	0.1521	0.1614	0.1578	0.1585	0.0057	0.0353
b_{10}	−0.0067	0.0088	−0.0052	0.0088	0.0015	0.1705
R^2	0.2018		0.2043		0.0025	

	Dirty		Clean		Difference	
Coeff	Est	SE	Est	SE	Est	St'dized
F Portugal						
b_1	−0.0482*	0.0018	−0.0620*	0.0060	−0.0138	−7.6667
b_2	−0.0260*	0.0024	−0.0279*	0.0024	−0.0019	−0.7917
b_3	−0.2061*	0.0479	−0.2123*	0.0483	−0.0062	−0.1294
b_4	0.0060*	0.0029	0.0062*	0.0029	0.0002	0.0690
b_5	−0.0292	0.0956	−0.1720*	0.0843	−0.1428	−1.4937
b_6	Not in the equation		0.0129*	0.0061	—	—
b_7	0.0906	0.1229	Not in the equation		—	—
b_8	−0.0098	0.0061	Not in the equation		—	—
b_9	Not in the equation		−0.0861	0.1247	—	—
b_{10}	−0.0052	0.0046	0.0160*	0.0074	0.0108	2.3478
R^2	0.2216		0.2270		0.0054	

9 Summary and Conclusions

This assessment of WFS data editing began with a review of the policies which were adopted, the reasons behind them, and the various issues which surround editing. A strategy for measuring the effects of editing was then developed. This strategy was limited by the unavailability of intermediate files from participating countries and also by the difficulty of distinguishing structural edits from other kinds of edits with full confidence. Despite these and other technical problems, enough early files and country-specific computer programs were located in the WFS archives to permit some systematic reconstructions. The data analysis began with some detailed case studies of raw data files, concentrating on the Ghana Fertility Survey. At this level we saw in the greatest detail the typical patterns of code changes which are made during editing. Thus, most of the out-of-range codes are simple data entry errors, such as column shifts, which could actually be corrected with considerable (although not perfect) accuracy simply from a knowledge of characteristic data entry errors. Other within-range shifts, although numerous, typically involve short distances.

In terms of practical implications, the most important part of the analysis dealt with comparisons of six dirty and clean pairs of files in diagnostic marginal distributions, two-way tables, fertility rates, and multivariate analyses. The dirty file in each pair was presumed free of structural errors but unedited for skip, range, filter and consistency errors. Perhaps the most notable weakness in the dirty estimates was that the fertility rates were too low. This deficiency has no connection with the editing, and was traced to a feature of the FERTRATE program which could be overcome with a programming change. Otherwise, even the rather elaborate logit regression and multiple regression in section 8 differ surprisingly little between the dirty and clean files. The multiple regression of children born in the last five years upon marital duration, age at marriage, type of place of residence, and education, plus interaction terms including duration, was particularly weighted towards variables that are subject to editing. Even so, both of the multivariate analyses were relatively insensitive to the editing. Specifically, the changes in inferences about the magnitude of effects and their statistical significance are almost always less than the differences that would exist between two independent and clean samples. There is also no evidence that dirty estimates are seriously biased or attenuated. These multivariate analyses were prepared with standard software, SPSS and GLIM, the only concession to the dirty files being the use of 'SELECT IF' statements to exclude out-of-range codes.

This report has also briefly reviewed the cost of machine editing, particularly in terms of the additional elapsed time and the depreciated value of the data which arises from such delays. We have rather crudely estimated that the average delay attributable to range, skip, filter and consistency checks was approximately one year, and the average cost was at least $75,000 per survey. Compared to the benefits, these delays and costs are unquestionably excessive.

In section 2 of this report we sketched some alternative strategies for editing. The first of these would terminate editing after a pre-set interval of time or expenditure. The second strategy would continue the editing until the number of known inconsistencies was brought below some pre-set level, and the remainder would then be put into a 'not stated' category. The third possibility would be to edit in stages, issuing a preliminary report, then a First Country Report, and then proceeding to second stage analyses, but spreading out the editing so that the reports which require less editing would not be delayed. Following the empirical work, these alternative strategies seem less pertinent, although for surveys with a larger volume of inconsistencies than WFS normally experienced, each alternative is worth considering.

One issue which does remain pertinent is the need for an early diagnosis of whether a country requires an extensive review of the questionnaires. There is not enough variation within our selected set of six countries to justify the development of an early indicator. It is possible that the initial proportion of out-of-range codes, for example, could serve this purpose. If this proportion exceeded some pre-set level, then one might decide to take the trouble to resolve inconsistencies by the laborious process of listing, consulting the questionnaires, and updating, until the point where a cost-specific, benefit-specific, or policy-specific level of such adjustments had been reached. We are not prepared to advocate either an indicator or a critical value which would serve this purpose. In any event, such a decision should apply to specific variables rather than to all of them.

We now turn to the main conclusions of this investigation. These will be phrased in the form of recommendations for any future surveys which are comparable, in a general sense, to those conducted by the World Fertility Survey.

The policies which directly concern machine editing are very closely connected with policies about other aspects of the total operation. It is in the development and application of the survey instrument that the true quality of the data is achieved. Thus, for example, the number of errors and the difficulty in correcting them will be reduced if the questionnaire has a simple structure. The WFS core questionnaire tended to invite some inconsistencies, often minor but difficult to resolve, by obtaining several event histories independently rather than in an integrated way; by repeatedly re-establishing basic filters, such as whether the woman is currently married or has ever used contraception or is currently pregnant; and so on. Of course,

the analytic importance of the data and the ease of the interview are other important concerns for the design of the questionnaire.

The training of interviewers and the care taken during field and office editing are believed to be critically important for the quality of the data. Their high standard of performance in most WFS surveys is almost certainly the reason why machine editing only detected a handful of inconsistencies in each case other than data entry errors.

The degree and manner of machine editing can be better co-ordinated with the analysis plan, a possibility which WFS only realized in its later surveys. A preliminary report, consisting of some key percentages, means, and even simple two-way tables, can be safely prepared from an early raw data file. A quick, preliminary report is particularly justified if it focuses on those variables which have little long-term interest and if it clearly indicates a tentative character. Then, even if an intensive editing policy is implemented later, users will have been provided with key figures which are probably within one or two per cent of their final values.

If data processing and editing are to be done in the countries themselves, and if this is an activity in which there is little local expertise, then it is critical that the local staff be trained, motivated, and organized to carry out the work rapidly.

The importance—whether psychological or more analytic —of agreement among subtotals, etc cannot be ignored. The structure of the codebook and of each individual case should follow the structure of the questionnaire. Especially for advanced and comparative analyses, it is indeed important that the data be clean. Users of the data expect them to be clean, and if they are not, users will embark on their own cleaning exercises. This will cause delays and inefficiencies of scale if several users work independently. One cannot deny the consensus among statisticians and demographers, as well as data processors, that data files should be free of inconsistencies.

For these reasons, we accept that it is indeed desirable to run edit checks, using software and specifications prepared in advance, and to achieve consistency in the data even though this consistency may have a negligible impact on the analysis. However, this study indicates that the machine editing should be done with greater efficiency. Our main recommendation is that the potential difficulty of achieving consistency should be anticipated, and a strategy should be developed well in advance, dependent in part on the circumstances. The process should be carefully monitored and not allowed to fall badly behind schedule. The fact that some WFS surveys were edited in only a few months clearly implies that the procedure described in section 2 can be conducted efficiently. It is possible that improvements in management and organization would have been sufficient to reduce the time interval dramatically in the other countries.

We suggest that another procedure should also be considered, either as an alternative or as a supplement to that described in section 2. This procedure, which shifts a larger part of the work to the computer, will be described very briefly. As the edit checks are made, a variable would be added to each case, giving the number of edit violations associated with the case, and a computer listing would give the number of violations of each specific type of check in the entire file. Only the questionnaires with the highest frequency of violations would be examined individually. These may involve column shifts which affect several adjacent fields, erroneous filters, etc. Those variables or specific types of checks which produce the greatest number of violations would also receive special attention; perhaps the specification was incorrect or a variable was consistently miscoded. As far as possible, discrepancies which remain after the review of a relatively small number of questionnaires would be resolved by two types of default rules: either change the code to 'not stated' or, in the case of multiple contingencies, give priority to one indicator over the others. These defaults can be implemented either in groups by a computer program or else on a case-by-case basis with updates, but do not require an unnecessarily time-consuming review of the questionnaires. Automatic defaults are not always easy to specify, yet it makes scarcely any difference how they are specified, as far as the analysis is concerned. It is the final, highest level of consistency—that which involves only a fraction of the discrepancies—which has the greatest potential for inefficiency. This level of consistency can be attained through almost arbitrary resolutions without any measurable impact on the analysis.

We have concluded that in several countries the World Fertility Survey invested too many resources and too much time in machine editing, and have suggested ways in which this activity could have been simplified. The ultimate source of delays, perhaps, was the incompatibility of certain WFS objectives, which were admirable individually but were difficult to integrate successfully. These included the desire to produce data of the highest quality, to do as much data processing and editing as possible in the countries themselves, and not to produce preliminary reports which might jeopardize final reports. In addition, WFS shared in a rather widespread illusion about the benefits of resolving all inconsistencies by reference to the questionnaires. With the advantage of hindsight, it now appears both that the costs of this approach were underestimated and that its benefits were exaggerated. It is unlikely that this conclusion could have been reached without the actual experience of the WFS programme.

References

Little, Roderick J.A. (1982). Sampling Errors of Fertility Rates from the WFS. *WFS Technical Bulletins* no 10.

O'Muircheartaigh, C.A. and A.M. Marckwardt (1981). An Assessment of the Reliability of WFS Data. *World Fertility Survey Conference 1980: Record of Proceedings* vol 3.

Rodríguez, Germán and John Cleland (1981). Socio-Economic Determinants of Marital Fertility in Twenty Countries: a Multivariate Analysis. *World Fertility Survey Conference 1980: Record of Proceedings* vol 2.

Appendix A—A Possible Monitoring Procedure

It has been mentioned several times that the data necessary for a systematic review of the costs and benefits of editing were not available. The following outline sketches the information which it would be useful to collect for this purpose. This is information which can only be collected as the editing is in progress, and almost on a daily basis during that time. It is proposed that the structural edit be included in the monitoring procedure; even though it is regarded as mandatory, its time and cost should be controlled.

I Phase of editing

A Structural edit
- 1 Preparation of specifications
- 2 Programming of specifications
- 3 In successive iterations,
  - a Check
  - b Reconcile
  - c Update

B Consistency edit
- 1 Installation of editing package
- 2 Preparation of specifications
- 3 Programming of specifications
- 4 In successive iterations,
  - a Check
  - b Reconcile
  - c Update

C Date edit
- 1 Preparation of extraction program
- 2 Installation of date edit program
- 3 In successive iterations,
  - a Check
  - b Reconcile
  - c Update

II Costs to be recorded at each of the above steps, including each step of each iteration

A Dates when begun and completed
B Number of days required by staff or consultants in headquarters
C Number of days required by staff or consultants in country, including pro-rated portion of travel costs
D Number of days required by in-country staff
E Computer time and costs
F Other related expenses

III Amount of editing: changes made at each step in I.B and I.C

A Number of changes by type (out of range, not stated, skip, filter, other inconsistency)
B Number of changes by whether a first correction or whether remaining from or created by an earlier correction
C Number of changes by whether or not the main variables are affected
D Number of changes which were resolved with or without the questionnaires
E Frequency distribution of the number of changes per questionnaire
F Save a copy of the tape at the end of I.A, I.B, and I.C (latter is already saved as final raw data file)