PN-AAN-568

ISN 31749

THE CONSEQUENCES OF SMALL RICE FARM MECHANIZATION PROJECT

Working Paper No. 53

DATA MANAGEMENT FOR ANALYZING THE CONSEQUENCES OF MECHANIZATION

By

J. A. Wicks and M. Sumiran
The International Rice Research Institute
Los Baños, Laguna

1981

The Consequences of Small Rice Farm Mechanization Project is supported by the United States Agency for International Development under contract tac-1466 and Grant No. 931-1026.01 and is being implemented by the International Rice Research Institute and the Agricultural Development Council, Inc.

ABSTRACT

Data management provides the link between the raw data set and the analytical process. With the complex sets collected in the Consequences project data management is not an easy process. In this paper we review the various basic requirements for effective data management and then examine the alternative approaches from raw data storage through to specific data handling packages. It is proposed that the best way to handle the data in order to satisfy project requirements is to simultaneously run FMDCAS and develop more complex systems.

INTRODUCTION

In a large and complex study such as the Consequences of Mechanization effective data management provides the key link between the data collection and the analytical stages. Within this context we define data management as commencing at the attainment of a "clean" data set and terminating at the achievement of the simple data file required for analysis. Many alternative data management strategies are possible, each of which will require different relative levels of programmer-analyst-computer capability. The purpose in this paper is to review some of the options available for storing the vast amounts of data collected from the four research sites, and to propose an appropriate strategy.

THE PROBLEM

Up to twenty nine different questionnaire forms were completed for each of the households included in the survey component of the study. Six different forms were used for farm recordkeeping. Additional data were also collected to help in generating some aggragate values. These can be stored for future use if desired.

The data collected may be summarized conveniently into a number of major areas, namely households, household members, land parcels, crop production, productive livestock, draft livestock, machinery, other assets and disposal of produce. The major areas may be considered as individual files and, if necessary, sub-areas may be specified within each of the major areas. By definition, each household on the survey provides one set of observations for the household file. However it is clear that each household may provide 0, 1 or many observations for the other files. This will depend, among other things, on household composition, whether it is a farm or landless household and, if a farm household, whether it is mechanized or non-mechanized.

Given the magnitude of the data collection effort, vast quantities of information have been gathered. If there are to made readily available to analysts, special care must be taken in defining data structures.

A range of options may be considered for this purpose. They are reviewed in detail in the following section.

BASIC CONSIDERATIONS

There are several approaches to data storage and management, the application of which depends on the availability and efficiency of the system. If a data management system is available, an evaluation should be conducted based on the objectives of the application. These objectives vary from application to application. Furthermore, the degree to which

each objective is given significance differs from an application to another. For research studies the constraints include:

Tools available. A major consideration in selecting a data management and analysis system is availability of sufficient facilities and tools for general data processing and analysis needs. Desirably, the system should have facilities for:

a) information storage and retrieval;
b) data modification
c) report writing;
d) analysis;
e) file handling.

The system should have programs and routines to accomplish these objectives available for use by researchers who will perform analysis on the data. Should a desired program not be available, the system should enable the researcher to develop his own program without knowledge of computing fundamentals.

Ease of use. Though the system may have good tools for data management and analysis, difficulty in using the system will prevent maximum utilization of its facilities. Ease-of-use is an important consideration in selecting a system. Although having a competent programmer may be able to provide a solution, it is not always possible, or beneficial, to have one available. It is usually better for a researcher to perform his own analysis, but this is only possible if the system chosen is easy to learn and use.

Performance and efficiency. It is expected that the data management system will produce data sets and output in a relatively short period of time. Often, research studies require immediate summaries and initial analysis of the information gathered. The system should be able to function effectively in establishing the data set and generate the output in the shortest possible time. This is good performance. Associated with good performance is efficiency or cost. Naturally, an economic system is desirable without sacrificing good performance.

Access flexibility. An important aspect of an effective data management system is its ability to access part of the data set on the basis of some criteria or classification. In many instances, researchers wish to limit analysis to some information or group of observations. In addition, they sometimes require a different format to enable transfer of data to another applications program. Further, requirements sometimes call for information to be extracted from different data sets. These and other needs call for flexible data access facilities.

Data modification, integrity and security. Different applications require different degrees of modification to the data set once it is

established. Not only may there be changes to specific information, but the creation of additional information from available data sometimes required. It is advantageous for the system to have facilities for which can be readily executed. This, however, brings up the problem of data security and integrity. Unauthorized or accidental modification must be avoided. The amount of modification done should be controlled and the system should contain facilities to control and record changes. Should a change be made on an observation, related data sets must also be checked to maintain a consistency.

Transferability. Some research studies require that data be used in several installations. If this is a major requirement, due consideration should be given to the availability of the system at several installations. Otherwise, data transfer would be difficult, if at all possible. This problem may be overcome by using a system which can be transferred from installation to installation without legal or system constraints. Alternatively, the system should either be one which is widely available or, at least, generate a set of data in a format accessible to different computer environments.

Documentation. For data management with a large number of expected users, documentation is a must. Problems arise when data items cannot be defined exactly. Keeping track of the contents of data sets is very difficult. Information often gets lost and this causes problems, for undocumented or badly documented files. It is, therefore, preferable to have a system which is self-documenting or requires researcher to document files and other data processing work. Work is easier for other researchers if data are properly and well documented.

APPROACHES

Using the above criteria, we can evaluate some of the common approaches to data management. It should be noted that the following approaches are not mutually exclusive. Frequently, they are overlapping when applied. The following breakdown has been so presented for simplicity.

Raw data storage. Data are transferred to computer-readable medium for later computer processing. This form constitute a raw data file. To retrieve data in this form and perform some analyses, program packages may be used without a package-defined data set being created (creating a package-defined data sets will be discussed later). With this approach, data are maintained as raw data and the researcher are free to select the package of programs he wants to use. He can develop his own program for specific applications. The tools and facilities available to the researcher varies with the programs to

which he has access and his expertise in developing his own programs. Ease-of-use will also depend on the exposure of the researcher to the program he proposes to use. Maintaining the raw data is however difficult. Each time an analysis is done, the researcher must define formats to read the data. Additional work is entailed with additional programming requirements for every analysis. Performance and efficiency is also low as program development takes a considerable amount of time. Working on raw data also poses difficulties in procedures requiring complex data access. Moreso, modifications on the raw data files is not straightforward. In addition, there is no guarantee to the integrity and security of these data files. Another disadvantage of this approach which often is overlooked at the early stages is documentation. Documentation is required to keep track of information on data items, variable codes, etc. On the other hand, the main advantage of keeping data in the raw form is transferability. For this one reason, it may be advantageous to maintain raw data files as a back-up while using another data management package for applications. Others use this as a starting point should there be problems in using more sophisticated data handling and analysis systems.

General Data Analysis Package

Systems consisting of programs and facilities for general data handling and analysis work may generally be as general data analysis packages. Examples are Statistical Analysis System (SAS), Statistical Package for Social Sciences (SPSS), and the Biomedical Computer Programs (BMDP). These provide many or all of the basic tools needed for data analysis, namely:

1) Data storage and retrieval;
2) Data access, computation and modification;
3) Programming;
4) File handling;
5) Data analysis;
6) Report writing.

With a good selection of a general handling package researchers are provided with the tools they will need. Because these packages are developed for general use, ease of use can be expected. Because of the many facilities and options offered to users by these packages use of programs from these systems may be costly. This is counteracted by savings in time and effort required to develop programs. Data and data sets access flexibility is dependent on the package selected. SAS, for example, is very flexible in accessing data sets and data to generate new data sets. Data modifications are easy to execute. Documentation is required to create package-defined data sets. Variables have to be defined and files need to be defined. Even then, researchers will have to utilize these facilities for his best benefit. Transferability is sometimes a problem. Data sets created and defined

for a particular package cannot be accessed, in general, in another installation if the system is not available. The trend shows, however, that more and more computer centers are installing these general data analysis packages.

Specific data handling program

An alternative approach to data management and analysis is through a specific data handling program. Programs of this type are input and output specific. Data collected should conform to the requirement of the system. Data formats should also follow the specifications of the program. Output from this program are of a specific format. An example of this type of system is the Farm Management Data Collection and Analysis System (FMDCAS). The tools provided by this type of systems vary. Some will have very limited, others will have sufficient routines. Researchers must therefore know the facilities they have before plunging into using the system. These systems are easy to use if all requirements are satisfied at data collection and computerization. This may not be easy to achieve. Deviation from the specified rules of the system can create several problems. Adding or deleting information to that required by the system will pose a lot of problems. In general, adding new routines is difficult. Other problems encountered in this system concern data access and modifications. Almost always, additional programming will be required to access part of the data set or if a new data set format is required. However, if the requirements of the system are satisfied by the data gathered, and all of the research output required can be provided through the system, this will be the ideal system for that research project. Such a system can provide output in a short time and at less cost. It is easier to transfer the system for installation at other computer centers. Because of its specificity, documentation is not a problem.

PROPOSED STRATEGY

At this stage in the project the major emphasis should be on analysis rather than collection of additional data. Further the time limitations imposed suggest that one strategy should be adopted to provide a short-run set of analyses, whilst an alternative strategy should be adopted for the logn-run. The strategies for the survey and farm record-keeping data may differ since the objectives and potential uses of the data are vastly different. We now propose strategies for each of these systems.

Survey data

Short-run analysis of the survey data should concentrate on obtaining rapid answers to some of the critical mechanization issues,

even if at the cost of some data loss. Hence FMDCAS is suggested as the appropriate mechanism. Progress with modification of the Consequences-designed data sets is described in subsequent papers by Webster, and Wicks and Lim. Satisfactory progress is being made with these changes and we anticipate that a full data set for one season should soon be ready for analysis using FMDCAS. Once this has been completed some initial analyses can be accomplished.

The long-run strategy is much less clear. Data storage in the form of raw data is particularly attractive since zero loss of information is ensured, at least initially and programmers time for constructing the data files is minimized. However, this approach has several very serious limitations. First, the communication problems associated with coordinating data collection at four different sites have resulted in some inconsistencies in formats across sites. Similarly, some data codes have different meanings at different sites. Standardization of these would facilitate data analysis, particularly cross-site comparisons. Second, the number of codes which must be appreciated, prior to even a minimum analysis, present any researcher with a complex problem in addition to the direct problems of analysis. Third, the number of questionnaire schedules completed would necessitate a close knowledge of their structure prior to any analysis, and fourth, the low level of the data itself would necessitate a substantial computing ability to implement analyses.[1/]

Probably the most useful alternative strategy is to store data in a set of SAS structured files. The file structure, and merging, capabilities of SAS, together with the widespread use of IBM systems, make it a very suitable candidate for data storage. The SAS data base stores data sets, variable names and variable labels, if required. These may be accessed for any run with the result that variable names and labels do not have to be respecified. Value labels are more complex to handle since SAS does not have the direct value label indicators that are present with SPSS. Three options are available for overcoming this problem. First either a new variable containing the value label could be defined. Second, the codes used for the actual variables could be replaced by value labels. Third, data sets for each of the codes and their respective value labels could be created and these merged as required for analyses. Of these options, the third appears to be the most preferable. It has the advantage that the code numbers are retained and thus permits quicker aggregation, and for cases where a set of value labels is used for more than one variable, only one set needs to be entered. For crops and rice varieties this would prove an extremely efficient approach.

[1/] This is evidenced by the current tendency of consequences of project researchers to extract data from the original coding forms.

In the light of this proposal we now turn attention briefly to the file structure which would be required for maintaining the survey data set. A possible schematic structure for the major files which could be maintained using SAS is shown in Figure 1. For ease of presentation, the figure shows a simplification, of a scheme that was reviewed earlier with individual site supervisors. Several minor files would need to be specified, as well as value label files but these are omitted here for clarity.

Although it is possible to start from the existing raw data files, a considerable amount of programming will be required to define the structures of each of the files. Also it will be necessary to ensure standard value labels and definitions across sites. Much of the initial preparatory work has been completed, and the major remaining constraint is the programming needs.

If such a system can be established, the benefits appear to be very considerable. First the number of variables which need to be understood have been significantly reduced. Second, the data can be accessed using standard computer routines and hence should not require large amounts of highly skilled programmer's time during the analytical stages. Third, different files can be merged readily to obtain a single data set for any analysis. The major disadvantage of the system is the initial input of resources required to establish the data sets.

Farm Recordkeeping (FRK) Data

Handling of FRK data requires an alternative approach since the format of the data collection activities was different, and potential users are directed more towards case studies than statistical exercises. Hence there are no benefits from summarising the many data items into aggregated values. However, ready access through a standard statistical package would doubtless encourage more widespread use. It is likely that this can best be achieved by storing the data from each form as an SAS file, with possibly some breakdown of forms such as Resource Utilization. Although it might be useful to some persons, inclusion of value labels would probably be too complex unless they can be obtained directly from the survey data sets. Hence the major impetus will be on the user to understand the data set.

SUMMARY AND CONCLUSION

Some of the important considerations for handling the complex sets of data collected in the Consequences of Mechanization Study have been reviewed in this paper. A number of potential techniques for handling the data have been considered.

Given the over-riding need at this stage to produce results, a two-pronged strategy has been suggested for the main survey data. The short-run approach should be to set up data sets using FMDCAS, even if there is a significant amount of information loss, and the longer-run objective should be to capture all facets of the data in a more comprehensive, yet easily used, system of data sets. For the FRK data, it is suggested that work should not proceed beyond specification of simple data files which can be accessed by users.

REFERENCES

Webster, J.P.G. (1981). "An Evaluation of Mechanization Data Using the FAO Farm Management Data Collection and Analysis System (FMDCAS)". Paper to be presented at a workshop on Consequences of Small Rice Farm Mechanization in Asia. IRRI. September 14-18.

Wicks, J.A. and P. Lim. (1989). "Future Plans for Philippines." Paper to be presented at a workshop on Consequences of Small Rice Farm Mechanization in Asia. IRRI. September 14-18.

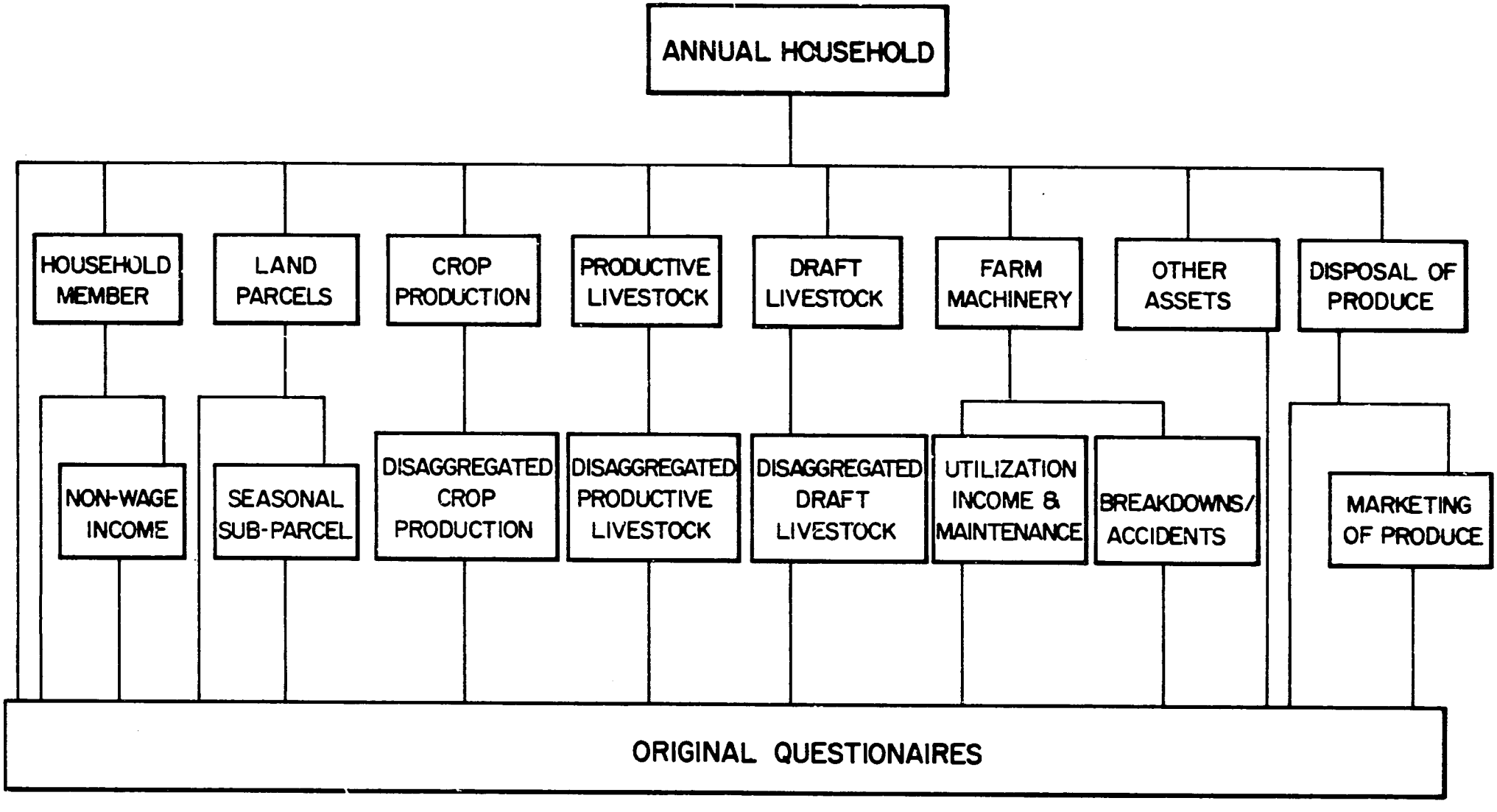

Fig. 1. Schematic representation of structure of major files.

CONSEQUENCES OF SMALL RICE FARM MECHANIZATION PROJECT

Working Papers

1. Juarez, F. and B. Duff. The Economic and Institutional Impact of Mechanical Threshing in Iloilo and Laguna. October 1979.

2. Pathnopas, R. The Economics of Rice Threshing Machines in Thailand: A Case Study of Chachoengsao and Supanburi Provinces. October 1979.

3. Gardezi, J., A. Rauf, M. Munir, K. Altaf, Q. Mohd-ud-Din, and B. Lockwood. A Study of Mechanical and Traditional Wheat Threshing in Multan District, Punjab, Pakistan: Some Preliminary Results. October 1979.

4. Habito, C. and B. Duff. A Simulation Model to Evaluate Mechanization of Rice Postharvest Operations in the Philippines. October 1979.

5. Chapman, J. The Potential of Mechanization for Crop Intensification in a Rainfed Area - Iloilo, Philippines. October 1979.

6. Thapa, G. The Economics of Tractor Ownership and Use in the Nepal, Terai. October 1979.

7. Jongsuwat, N. Productivity Growth and Farm Machinery Adoption in Thai Agriculture. April 1980.

8. Bernsten, R.H. and R. Sinaga. A Methodology for Identifying Lowland Rice Farms that Would Benefit from the Mechanization of Land Preparation. October 1979.

9. Bernsten, R. H. and A. Rochim. Labor Shortage as a Constraint to Increasing Cropping Intensity. Revised March 1980.

10. Ayob, A.M. The Economics and Adoption of the Combine Harvester in the Muda Region of Malaysia. October 1979.

11. Lubis, R. Impact of Cropping Pattern Technology on Income, Employment and Production: A Case Study on Expanded Crop Production in Lampung. October 1979.

12. Wicks, J. A. Modelling the Consequences of Future Mechanization: An Outline of Possible Procedures. October 1979.

13. Khoju, M.R. and J. A. Wicks. Economics of Pump-Irrigation in Eastern Nepal. August 1980.

14. Tan, Y. L., J.P.G. Webster and J. A. Wicks. The Decomposition of Differences in Output Between Two Groups of Farms. Revised 1981.

15. Herdt, R. W. Mechanization of Rice Production in Developing Asian Countries: Perspective, Evidence, and Issues. September 1981.

16. Lantin, R. M. Mechanization Policy and the National Agricultural Mechanization Council - Philippines. September 1981.

17. Lockwood, B. Farm Mechanization in Pakistan: Policy and Practice September 1981.

18. Wiboonchutikula, P. The Total Factor Productivity Growth of the Three Digit Manufacturing Industries in Thailand. September 1981.

19. Ahammed, C. S. and R. W. Herdt. A General Equilibrium Analysis of the Effects of Rice Farm Mechanization in the Philippines. September 1981.

20. Mikkelsen, K. and N. Langam. Technology Change in the Philippine Agricultural Machinery Industry. September 1981.

21. Boughton, D. Energy Use in Alternative Rice Production Systems in Nueva Ecija, Central Luzon, Philippines. September 1981.

22. Hurun, A. Financial Analysis of Power Tiller Ownership in Mariuk Village, West Java, Indonesia. September 1981.

23. Hafsah, J. The Economics of Tractor Operation and Use in South Sulawesi, Indonesia. September 1981.

24. Maranan, C. L. A Comparative Analysis of Tractor Contract Operations in Nueva Ecija, Philippines, 1972 and 1980. September 1981.

25. Monge, V. S. and B. Duff. Analysis of the Demand for Farm Power for Small Rice Farm Agriculture in Nueva Ecija, Philippines. September 1981.

26. Munir, M. An Evaluation of the Farmers' Decision-Making for Investment in Farm Machinery. September 1981.

27. Jabbar, M. A., Md.S. R. Bhuiyan and A. K. Maksudul Bari. Causes and Consequences of Power Tiller Utilization in Two Areas of Bangladesh. September 1981.

28. Juarez, F. and R. Pathnopas. A Comparative Analysis of Thresher Adoption and Use in Thailand and the Philippines. September 1981.

29. Ahmed, J. U. Labour Use Pattern & Mechanization of Paddy Postharvest Processing in Bangladesh. September 1981.

30. Colter, J. M. The Impact of Handtractors on Income and Employment Opportunities of Migrant Laborers in Java. September 1981.

12

31. Santoso, K. The Potential for Agricultural Mechanization and Labor Markets in East Java. September 1981.

32. Bernsten, R. H. Effects of Mini-Tractor Mechanization on Employment and Labour Use Intensity, Sidrap and Pinrang, South Sulawesi, Indonesia. September 1981.

33. Collier, W. Improved Cropping Patterns, Labor Absorption and Small Farm Mechanization in Indonesia. September 1981.

34. Moran, P. B. and E. Camacho. Consequences of Farm Mechanization Project Site Description: Philippines. September 1981.

35. Generalla, A. C. and A. Aguilar. Effects of Mechanization on Intensity of Land Use. September 1981.

36. Tan, Y. and J. A. Wicks. Production Effects of Mechanization. September 1981.

37. Sison, J. F. and P. B. Moran. Farm Labor Utilization and Employment in Two Selected Municipalities in Nueva Ecija - A Preliminary Analysis. September 1981.

38. Saefuddin, Y. Site Description: Mechanization Consequences Project in West Java, Indonesia. September 1981.

39. Handaka, S. Effects of Mechanization on Intensity of Land Use, West Java, Indonesia. September 1981.

40. Handaka, S. A Technical and Economic Evaluation of Rice Mills in West Java, Indonesia. September 1981.

41. Sinaga, R. Effects of Mechanization on Productivity: West Java, Indonesia. September 1981.

42. Sinaga, R. Effects of Mechanization on Productivity: South Sulawesi, Indonesia. September 1981.

43. Sri-Bagyo, A. The Impact of Mechanization on Production and Employment in Rice Areas of West Java, Indonesia. September 1981.

44. Maamun, Y. Site Description: Mechanization Consequences Project in South Sulawesi. September 1981.

45. Sarasutha, I. G. P. and R. Bernsten. Effect of Mechanization on Intensity of Land Use, South Sulawesi, Indonesia. September 1981.

46. Bockhop, C. W. and M. Nafziger. The Impact of Economics upon the Design of Machinery at IRRI. September 1981.

47. Wattanutchariya, S. Economic Analysis of Farm Machinery Industry and Tractor Contractor Business in Thailand. September 1981.

48. Hussain, K. A. An Assessment of Capacity of Workshops and Farmers To Repair and Maintain Farm Machinery in District Faisalabad: Summary of Major Findings and Policy Recommendations. September 1981.

49. Gonzales, L. A. and R. W. Herdt. Evaluating the Sectoral Impact of Mechanization on Employment and Rice Production in the Philippines: A Simulation Analysis. September 1981.

50. Khoju, M. R. The Economics of Pump Irrigation in Eastern Nepal. September 1981.

51. Sudaryanto, T. The Effect of Tubewell on Income and Employment: A Case Study in Three Villages in Kediri, East Java, Indonesia. September 1981.

52. Santoso, K. Economics of Pumpsets in East Java. September 1981.

53. Wicks, J. A. and M. A. Sumiran. Data Management for Analyzing the Consequences of Mechanization. September 1981.

54. Webster, J. P. G. An Evaluation of Mechanization Data Using the FAO's Management Data Collection and Analysis System (FMDCAS). September 1981.

55. Lingard, J. Measuring the Impact of Mechanization on Output. September 1981.

JW/hhr