Lanny Norensberg

I am here to speak on SPSS. Many people have called me to say that they would be here to listen. SPSS is not up yet, however, we are looking forward to getting something up in the near future. Unfortunately, it is a slow process. As you know, SPSS is now up on various 360's, 370's, CDC, PDP 10's. and other machines. There is a "version" on the 2000 written in Basic, but it is not a version sanctioned by SPSS. As of now, there is no version on the 3000.

About two years ago, in response to an internal need that we at NERA had, I contacted SPSS to see what could be done to implement the system on the HP3000. They said the only thing they could do is send me a tape and a McGraw-Hill book and their best wishes. And, that is what they did. They sent me a tape with 93,000 source cards.

The problems with SPSS are as follows. Before FORTRAN B, much of the task was impossible. It could not be done without spending large sums of manpower, money, time, etc. The problems are as follows. Before FORTRAN, you did not have options such as entry; if that option was needed, and SPSS has used it considerably, you had to rewrite the subroutine. And, considering that there are several hundred subroutines in that package, this was a mammoth task. Also, you never know if they just stay at that entry point or whether they actually go back to the beginning of the subroutine. You have to trace everything down. We were spending a great deal of time on it. Also, it must be realized that we cannot devote six people to this effort full-time. We are a profit organization. My computer department services the economists at NERA. SPSS is just a sideline. We initially went into it to serve ourselves. As it turns out, we may end up servicing the entire Hewlett-Packard community.

Some of you may not know what SPSS is. It stand sofr "Statistical Package for the Social Sciences." SPSS was developed by the National Opinion Research Organization at the University of Chicago. They developed it for their own internal needs and found out that it was very marketable. However, as I pointed out at the beginning, it was marketed for the IBM360. Eventually, they moved it up to the IBM370 and then eventually others converted it to the other machines.

Only one of the problems with SPSS was pointed out. There are several others. For instance, many of the key routines were written in BAL, which is IBM's assembly language, which makes it impossible to run those routines on our system. We either have to convert them or forget about them. Where we can, we are choosing to forget about them, at least for the moment. Where we are able, we hope to convert them. The only problem here is to find out what those routines do. SPSS, on many long calls to them, and in some letters to them, has been unable to tell us completely what those routines do.

Now that we have FORTRAN B, some of the problems I mentioned before, such as entry, double integer and a few others have been solved. But, not all the problems have been solved. For instance, they use LOGICAL *1 quite often. The question arises as to whether logical one is used as a character or is used as a logical. I have found both instances in subroutines. The same variable has been used as a byte character and as a logical. The question is how to convert it. Do we create new variables, have two variables running down the line, etc?

There is another problem with SPSS. Apparently, at least, our compiler has been flagging it. When on subroutine calls another subroutine, it passes the variable either as real, double precision, integer, etc., in the same dummy position. They will call a routine five or six times all with different parameters. The FORTRAN B compiler flags this. FORTRAN B, as soon as it gets the first subroutine call, looks at the parameter list, keeps a record of it, and then when it gets the second subroutine call it says, "Something is wrong." This subroutine call does not match the first one, and flags it. This means I have to go back to the original routine and find out why it wants it in one of these formats and rewrite the subroutine call to make sure it gets the data that it wants at that precise moment. It is not very simple, especially when you have several hundred routines and any 80 of them at one time may be calling the same subroutine. It is a huge problem.

Over the summer, I had two and one-half people work on it full-time for two months just trying to figure out which subroutine called which subroutine. We are finally getting to the point where we may lock that problem. Now we have to resolve it.

These are only the minor problems. Then major ones start. The version we have is version six, which is 156K version that is on 360. We should have no problem fitting in the same size data base simply taking advantage of the virtual memory aspect of the 3000. They rely totally on the whole thing residing incore. We can segment the program a little better than they can. But, one of the fears in segmenting it, is that we will oversegment. If it takes one-tenth of a second to call in a segment, if you oversegment it, especially if you are calling a subroutine a thousand times, you can end up running very slowly. So, that is one of the problems we have to solve.

I have requested that Hewlett-Packard provide some aid and assistance in this, only to be turned down. They are not a very helpful company, especially since they seem to be vending the SPSS package for me. I receive calls from around the country from people saying that a Hewlett-Packard salesman has told them that I have a finished product and that I am selling it. And, the Hewlett-Packard salesman seems to be selling the 3000 on the basis of my product. This I find very distasteful, considering the fact that when I call them up and ask for help they turn me down saying it is not one of their products; it is one of NERA's products and they do not want to be bothered with it. But they seem to be doing a very good job of marketing it.

What we are hoping for eventually is to come out with a small version of this package, not a full version. From our analysis, of not only our work at NERA, but from discussions with others dealing in economic research, I have discovered that there are two main programs everybody wants. One is a regression analysis--a super sophisticated one. And the other is a cross-tab with all its accompanying routines. The other things that SPSS has are nice and people would like them, and hopefully we will be able to get them up, but immediately we are striving to put up a stripped down version of regression and cross-tabs. If we are lucky, and everything goes our way, and the 3000 is cooperative and stays up, and the staff is still available to me, I hope to get something up sometime before April.

Admittedly, it will not be as neat to use in some cases as SPSS. We will have to take away some of those super fast read routines, at least temporarily, until we substitute them with SPL routines that can read a little faster. In the meantime, what we will do is stay with FORTRAN, get a nice, neat FORTRAN version up that will read your data, that will read your control cards, strip them down, analyze them, etc. I do not guarantee that it will run fast. For those of you that are using it in a high production shop, or intend to use it in a high production shop where you must meet instantaneous deadlines, I do not believe the first version will be for you. In fact, it may not even be for us. We have an interest to speed it up. The first version will just simply produce accurate regressions.

We are going to have to spend much time augmenting the system. Every subroutine has to be gone through with a fine tooth comb to figure out exactly what it does. Then we must take out the unnecessary things. These are the problems we are going to face with SPSS.

It is now September 30. Tomorrow is October 1. Some of you have talked to me over the past six or seven months and always hear a new deadline pushed out in the future. But, consider the complexities of it and consider the fact that we keep finding more problems. Every time we correct one bug, we recompile it only to find out that the compiler missed several other bugs along the way.

There is another problem we have. They use hexadecimal throughout. They use it in data statements. FORTRAN B flags hexadecimal in data statement. A core-to-core conversion or reading it from some dummy file must be used just to produce the same results they do. They have character strings throughout. We have to make sure we are reading the same amount of characters they are reading.

I have tried to trace down several routines myself only to get lost in what the main control cards do and where they go. The manual is simply the same manual you can buy in any bookstore. The documentation SPSS has supplied is nil. All they have are quite a few comments in each program, which are good. But, as you can see we are running slow.

In the meantime, what NERA has done (if any of you want to send us a letter, we will be glad to make some of these available to you) is to develop some of our own internal routines to compensate. We have developed our own internal regression programs that do much of what SPSS does but without some of the output, and without the same need control cards. We have also developed our own cross-tab package. We had developed these packages before we contacted SPSS. We just polished them up a little bit better since then. If anyone is interested, feel free to write to me at National Economic Research Associates, Inc., 80 Broad Street, New York, New York 10004.

SPSS is not the only package we viewed. For instance, we looked at IDA. At Rockville, Maryland, we had converted IDA from 2000 BASIC to 3000 BASIC only to find out it was a "bomb." IDA did not serve our purpose. It works on small data sets, it was difficult to handle, no one wanted it. IDA is an interactive data analysis package. It is something they developed on a 2000. I think Chicago also did that. It is available at NYU's computer center; it is available on several others. You can get it for whatever the contributive library costs you. IDA is on one of the reels and there are about 30 or so BASIC subroutines. You can compile it. There are still problems. You have to figure out what IDA does. Just like SPSS. There is very little documentation on it. It is only a user's guide. But, it is simpler to put up than SPSS. For many of you who work on small data sets, that may be an advisable solution. Simply take the programs, compile it. The only thing is, of course, if you are going to compile it you have to figure out the entry points of each subroutine. And the entry points are not known in advance. You have to trace it down as we did. We do not have it completely worked out, but I will be glad to send you what we do have. Again, you can contact me for that and you can finish working on it yourself. If I ever get a working version, I will put it into the library.

Besides SPSS, we looked at ECON. We are converting ECON. As soon as we get it completely converted, we will put it into the library. It is a little simpler than SPSS. There are fewer routines, very little source code, and as soon as we finish it we will distribute it.

The big question is why did NERA get involved in all of this? We are a profit organization. We are not a university. We are not a public research firm. We are a private research firm. We work for utilities; we work for major corporations in antitrust. We are a firm of several dozen economic analysts who analyze a company's situation, not as a market study, but as a litigation situation, such as an antitrust or rate setting case.

We were finding out that many of the antitrust cases require several years of data over several companies. We were analyzing 10,000 observations, 20,000 observations--which may not seem like much to many other companies, but to companies like NERA, which was used to something like 60 or 70 observations at a time, it was an immense task.

Our economists found themselves using these large packages. Now we were obtaining the Hewlett-Packard 3000, a considerable investment. We had a decision to make. Do we continue going outside while we had this computer sitting inside or look to use it. Many people were using SPSS. They suggested we get it. We got it. Unfortunately, it is not in a state to help anybody.

I have discovered one thing in my life; there is no such thing as portability. IBM will prevent that. They go out of their way to prevent that. They do not care what the standards are; they come with IBM standards. Meanwhile, everyone else is developing along the other standard. Hewlett-Packard, however, should have come out with some of these standard items a long time ago. Entry, for instance, has been standard for a long time. Why it had to wait for FORTRAN B, I do not know. We converted a spline program which we received from some university; and fortunately it was a small one. It had entry all over the place. I had to resubroutine the thing.

You must realize that SPSS was specifically written for a 360 and it took advantage of whatever features IBM had. For instance, routines were written in BAL. You cannot expect Hewlett-Packard to convert to BAL. However, you can expect Hewlett-Packard to implement LOGICAL *1. I can expect them and do expect them to implement the hexadecimal data statement. That is a problem. We get many packages like that. RAPE has that, TSP has that. I could go through a whole list where they wrote BAL routines. What are we going to do? Perhaps, if Hewlett-Packard were a nice company, which lately they have shown otherwise, they would sit down and write a little assembler or something that would transfer BAL to SPL. Easy enough. They have much manpower out there. Let them do it. They want to sell these packages. They want to sell their machines. Other conversion programs have been written. I, myself, would undertake it if I knew BAL, but I do not know BAL well enough.

IMSL provides a package of subroutines. When I was at the Federal Reserve, we had purchased that package of subroutines. Our business is not to market SPSS. We are putting it up for ourselves. If it happens that it works, we will market it. We are not out to market it, per se. We are not going to make a profit on this thing. Most of the people who want it are universities and their rates are low. We are doing it for ourselves. IMSL is a good package. Remember, they have also written them for IBM systems. You will have to convert them. We had to convert them for a Burroughs 6700. And, a Burroughs 6700 is a much more powerful machine than any of the IBM machines that are used for the ISML package. There are many other packages out there. For those of you that have needs that are immediate, there is nothing wrong in going into the scientific subroutine package now that it is totally converted. You have the 1130 subroutine package which is an SL on the system. You have all these things. There is nothing wrong in writing your own front ends.

People are reluctant in writing their own programs these days. It is the calculator generation. No one knows how to add one plus one. They know how to press one, enter, plus one. That is all they know how to do, but they cannot add one and one on a piece of paper. One of the things that you have to learn is that if you do not have a package, do it yourself. That has been my philosophy. Do it yourself. You lift up a pencil and paper, you write out the code and you have a front end. The subroutines are available all over the place. They are public domain. They are free. You do not even have to pay exorbitant fees for them. Some are good; some are bad. You throw away the bad ones. If I came out and said SPSS was up, alive and working, and I guarantee it, you would accept it as is? When we received our 3000, I had a staff sitting down for days testing out every bit of a FORTRAN compiler to see if every single part of it worked according to specifications. And, it did not. I learned a long time ago not to believe anybody. You mean you are going to believe manufacturers? If you are interested in a regression package, you can get the Wampler and Longley test data. It is very good test data. Immediately, you can tell if the programs are bad. At the Federal Reserve, I had someone sit down and make up phony data and sit there and calculate the things by hand. The whole staff calculated it out to twenty-three decimal placed by hand. And then we ran the data through our regression programs to get the exact same results.

As I was saying before, just as a quick summary, SPSS is a very difficult package to convert. I may be somewhat masochistic in doing it. I may also be more than that, but, I am not making any total definite promises. May I suggest that nobody buy a 3000 on the basis that I will have SPSS up. I can think of many good reasons to buy a 3000 for those of you who are potential customers. It happens to be what with all my kidding (there are many faults in the system) better than most choices you can make. But, if your sole purpose for buying the 3000 is because I am going to have SPSS up, may I suggest that you forget it.

I hope to have SPSS up. I am going to try. We are working on it. If we can get something up in the near future, you will have it. But, as you can see with all the problems I am having (I probably have not even encountered half the problems yet) because as soon as I get a clean compile, I find out there are a thousand other problems in the system. So who knows when we will solve that. If you do want some of the packages we have developed, the regression package, the cross-tab package, contact me.