

195

VALIDATION STUDY PHASE I
FINAL REPORT

Prepared by:

Linda K. Holt
EDP Programmer III

Massachusetts Department of Correction

William T. Hogan, Jr.
Commissioner

June, 1980

ABSTRACT

This report presents the results of the first phase of a validation study. In this report information in the offender-based Correction and Parole Management Information System (CAPMIS) was compared to official documents in offender folders for 16 different variables representing personal characteristics and prior criminal history. Discrepancies were found to exist between the data base and folder information. The variables studied ranged from perfect agreement to forty percent discrepancy between the two data sources.

INTRODUCTION

In doing survey research, a concern of every social scientist must be the quality of the data that is used. An important aspect of data quality is its validity or how well the data represents what was supposed to be measured.

The issue of data validity is particularly critical in the case of the Department of Correction's data. This data forms the basis for a large number of policy studies. Any bias or inaccuracy could have serious consequences for policy makers and invalidate research efforts. Of equal importance is the unavailability of data due to the limited number of variables on which information is presently collected.

The purpose of this study is to assess the validity of the offender-based information system used at the Massachusetts Department of Correction and to determine the feasibility of adding new variables to it. The first phase of this study is reported here. It involves comparing information about an offender contained in the Correction and Parole Management Information System (CAPMIS) with official documents found in inmate folders. The results of phase two and three will follow in separate reports. In phase two inmate self report will be compared with information contained in CAPMIS. In phase three the possibilities of expanding the current data base will be explored.

The use of official documents for research in criminal justice has been questioned before. Official reports of criminal behavior may only represent a portion of actual behavior. Self reports of both apprehended and non-apprehended individuals indicate that people in both groups have engaged in more types of criminal activities than are indicated by official records (Hindelang, Hirschi and Weis, 1979).

Furthermore, information in inmate files often have been found to contain inaccurate or even contradictory information (Singer and Gottfredson, 1973).

There is some evidence that bias is likely to occur in an interview when there is social distance between interviewer and respondent. "Status distance and threatening questions may create a situation in which the respondent feels pressure to answer in the direction he believes will conform to the opinions or expectations of the interviewer" (Williams, 1968). This problem becomes particularly acute in corrections where information is used to decide placement, parole conditions, research activities, and other issues that directly affect an offender's experience within the corrections system.

There are contradictory theories about the possibilities for accuracy when measuring behavior and background characteristics of offenders and other deviant groups. On the one hand there are some who believe that such populations are motivated to conceal, and will not reveal accurate information (Becker, 1963). On the other hand, some believe that given the proper circumstances accurate data can be retrieved (Kinsey, 1948). In a study of drug addicts, Ball (1967) found data to be quite accurate. His study compared official data from a variety of sources with an inmate's self reports. However, accuracy was dependent on the way in which the data was collected. Comparisons between a variety of data sources should indicate the accuracy of Department of Corrections data.

Most of the data used at the Department of Corrections can be thought of in terms of standard survey research where data is collected through interviews. There are several important differences though. The interviews are done by police, probation officers, correctional counselors and parole officers rather than by persons trained in

interviewing for social research purposes.

There is no standard interview schedule that is followed. Information is gathered in certain broad areas as seems pertinent to the particular case and individual. Information on any variable can be found in depth for some cases while it may be non-existent for others. The data collection process is beyond the control of the research staff for the most part. However, this does not mean that an assessment can not be made of data accuracy. Improvements in accuracy may also be possible.

Offenders, like any respondent in a social survey, may also prove to be a source of inaccuracy of the data. This is always a concern because respondents are always free to manipulate the information given. This becomes more problematic when that information is not given anonymously and is often incriminating.

Data validity problems come from a variety of sources. Offenders may be motivated to manipulate the information they provide in order to assure an easier or shorter term of incarceration. The social situation in which information is collected may also lead to invalid data. Other standard sources of error including coding and keypunching may also affect data validity although more randomly. The particular mode of data collection, although it approximates survey research techniques, introduces other problems. This study should help to locate some areas of validity that are problematic as well as to alleviate questions about validity in other areas.

METHODOLOGY

A ten percent random sample of the 999 commitments to the Department of Correction in 1978 was drawn yielding a sample of 100 cases. For each person in this sample a listing was made of the values recorded on the computerized data base for each of the variables under consideration.

The variables that were studied included both personal background characteristics and criminal history. The 16 variables studied are:

- Military Service
- Marital Status
- Race
- Date of Birth
- Last Grade Completed
- Special Education
- Time on Job of Longest Duration
- Drug Use
- State Incarcerations
- County Houses of Correction Incarcerations
- Juvenile Incarcerations
- Adult Paroles
- Juvenile Paroles
- Parole Violator Commitment Date
- Date of First Arrest
- Parole Eligibility Date

The department folders of this sample were obtained and for each of the 16 variables an attempt was made to find any three documents in the folder that had information regarding the variable in question. From these three sources a decision was made regarding the appropriate value for that individual on a variable.

Since this is a commitment file, information should reflect a person's status at the time of commitment. It is quite likely that several variables may change as a result of a commitment (marital status and education for example), but the file should reflect the original status.

In many cases information from various sources was contradictory. In order to select from among the various sources of information the best one, a set of criteria for selecting the most accurate response was established.

A piece of information will be considered valid for the purpose of this study based on the following five criteria:

1. Date of document from which data was taken is near the date of commitment. Whether a document was prepared before or after the actual date of commitment, it will be considered more valid if the information was collected near the commitment date rather than long before or long after.
2. Verification of data by checking one source against another. In any case where attempts are made to verify data and such verification is noted by supporting documentation, a piece of information will be considered more valid.
3. Details of a document. Any account of a variable that has detailed information accompanying it will be considered more valid. Notes, narrative accounts, explanatory details will be considered to add validity to the source.
4. Number of documents. When a large number of documents attest to a particular value of a variable then more validity will be given to that position than to a minority view that is presented.
5. Legibility and originality. Since many documents are old, copied and in poor condition, those in better condition or original copies when available will be considered more valid.

Using these criteria, a summary assessment was made regarding whether there existed a discrepancy between the data base and the subject folder or whether there was agreement between the two sources. The cases where there was discrepancy are summarized and further studied to locate where and how those discrepancies occurred. When there is a pattern to those discrepancies it will be noted to ascertain whether there is a perceptible bias in the data base or the folder information.

Findings will be presented in three main categories: cases where information is missing in both the data base and offender folders, cases where there is no discrepancy between the two sources of information and cases where there is a discrepancy between the data base and folder information. The final group will be broken down by the type of discrepancy that occurred. Because the sample size was 100, percentages will not be presented. Findings will also include a summary of all variables to indicate their relative position.

FINDINGS

1. Military Service / Discharge

Fifteen percent of the sample contained disagreement between the computer data base and the offender's folder on this variable. Two-thirds of the disagreements involved cases where the data base indicated missing information and the folder supported either the presence or absence of a military career. In another case a discharge listed as unknown could be interpreted. In four cases individuals were listed as having no military career while their folders indicated some time in the service. In all cases the information could be found in probation or classification reports as well as other supporting documentation. It is not known to what extent those cases which do agree are actually valid. There is reason to believe that a bias exists toward an underestimate of undesirable discharges since 5 cases would be undesirable according to the folder and only 1 would be honorable.

VALIDATION STUDY RESULTS MILITARY SERVICE/DISCHARGE

DATA BASE AND FOLDER COMPARISON

	<u>N</u>
Missing and Remains Unknown	0
No Discrepancy	85
Discrepancy	15

<u>Data Base</u>	<u>Folder</u>	
Missing	No Service	9
Missing	Honorable	1
No Service	Undesireable	3
No Service	Honorable	1
Discharge Unknown	Undesireable	1

Total	100
-------	-----

2. Last Grade Completed

There were 27 cases with discrepancies between the data base and folder information regarding the last grade completed.

In almost half of the cases (13) the discrepancy was of one year only. This is usually the result of ambiguity about when a person left school and whether a grade was entered or completed. In five cases the data base indicated that the information was missing while the folder indicated an educational level for the person. In four cases the folder indicated a person received a GED prior to the date of commitment while the data base recorded only the formal educational level. In five cases the data base and folder information were discrepant by several years with no apparent explanation. In four of these five cases the data base indicated that the individual had 3 to 5 years more education than folder information indicated.

VALIDATION STUDY RESULTS
LAST GRADE COMPLETED

<u>DATA BASE AND FOLDER COMPARISON</u>		<u>N</u>
Missing and Remains Unknown		0
No Discrepancy		73
Discrepancy		27
Discrepancy of one year	13	
Missing data present in folder	5	
GED not counted in data base	4	
Data base more than folder	4	
Data base less than folder	1	
<u>TOTAL</u>		<u>100</u>

In general, probation reports and classification reports provided better information than a person's Quick Reference Index (QRI), a summary document providing encapsulated information about an offender. While the number of discrepant cases is greater than most other variables, the amount of discrepancy is not great. Since most discrepant cases are off by only a year in either direction the effect of these discrepancies should not be great.

3. Race

Of the one hundred cases sampled, there was discrepancy in two cases between the data base and folder information regarding race. In both of these cases the individual was hispanic according to the folder and black according to the data base. In both cases the folders had internal discrepancies. Some documents indicated that the person was hispanic while other documents indicated the person was white or black.

In the total sample of one hundred, there were four cases listed by the data base as hispanic. Considering only these four cases where there was agreement between the two data sources and the two cases where there were discrepancies, one-third of all hispanic cases involved discrepancies.

In another case while there was substantial agreement between the data base and folder information (indicating the person was white), some of the documents in the folder indicated the person was Portuguese. The ability to distinguish this group may be needed for some things.

4. Marital

Eight of one hundred cases had discrepancies between the computer data base and folders regarding marital status. Half of these cases (4) were recorded as single in the computer data base when the folders indicated past marriages ending in divorce. In three cases the data base indicated a person was married while their folder indicated a separation or divorce at the time of commitment. In the final case, the data base indicated marital status was missing while the folder indicated he was separated from a spouse.

VALIDATION STUDY RESULTS
MARITAL STATUS

<u>DATA BASE AND FOLDER COMPARISON</u>		<u>N</u>
Missing and Remains Unknown		0
No Discrepancy		92
Discrepancies		8
<u>Date Base</u>	<u>Folder</u>	
Single	Divorced	4
Married	Separated/Divorced	3
Missing	Separated	1
<hr/>		
TOTAL		100

5. Date of Birth

In seven of the one hundred cases sampled, there was disagreement between the computer data base and the folder information regarding a person's date of birth.

In four of these cases one digit of the six digit date of birth was in error resulting probably from a typographical error at some

stage in the building of the data base. For example the data base read 10/25/55 while folder information gave 10/20/55 as the date of birth. Two of these cases resulted in discrepancies of less than a year, two resulted in discrepancies of a year or more.

In the other three cases of discrepancies resulted from a variety of reasons. In one case a twenty-year discrepancy resulted perhaps from one document, the QRI, having a wrong date. In another case there was no apparent reason for a discrepancy of over a year. The final discrepancy came in a case where the individual's age is unknown and the data base recorded the latest possible birth date.

6. Date of First Arrest

While this variable is referred to as the date of first arrest, it would be more accurately referred to as the date of first court appearance since this is the date that is usually given and information that is readily available.

There were fifteen cases where there was a discrepancy between the data base and the information in a person's folder. In over half of these cases, 8, the data base did not reflect an early arrest that occurred in another state or as a juvenile that was not included in the probation report. In three cases the information in the folder does not reflect an earlier date given in the data base.

In three cases a Massachusetts adult charge was reflected in the folder that was not on the data base. A final case seems to be a typographical error in the month.

VALIDATION STUDY RESULTS
DATE OF FIRST ARREST

<u>DATA BASE AND FOLDER COMPARISON</u>	<u>N</u>
Missing and Remains Unknown	1
No Discrepancies	84
Discrepancies	15
Early Juvenile/Out of State Charge	8
Folder gives no mention of arrest	3
Early Mass. charge not included	3
Typographical error	1
<hr/>	
TOTAL	100

7. Time on Job of Longest Duration

There were 40 cases where folder information and the data base did not agree on the time a person spent in their longest job. In over a quarter of the cases (11) work history was listed as missing when a person had a work history. In some of these cases a person's work history was nominal or even non-existent but the folder had evidence supporting this.

In 16 cases the data base estimated the time on longest job as less than the person's folder. In the other 13 cases the data base estimated a longer time period than was documented in the folder.

VALIDATION STUDY
TIME ON JOB OF LONGEST DURATION

<u>DATA BASE AND FOLDER COMPARISON</u>	<u>N</u>
Missing and Remains Unknown	2
No Discrepancy	58
Discrepancies	40
Missing Data Available	11
Data Base Less Than Folder	16
Data Base Longer than Folder	13
<hr/>	
TOTAL	100

Not only were the number of cases with discrepancies large but the change in the coding would have been great in many cases. Since the codes represent varying lengths of time, the amount of discrepancy was determined by subtracting the codes. In order to bring the data base and folders into agreement, the average change would have been 2 codes (i.e., a 1 to a 3 or a 5 to a 7, etc.).

This variable had a large number of discrepancies for the women in the sample. Five of the seven women in the sample had discrepancies between the data base and folder information. In four of these five cases the data base indicated less work history than the folders.

Discrepancy may come from a number of sources. Some probation reports are prepared quite a while prior to the time of incarceration, and individuals keep working up until the time of commitment. Many documents do not record dates, or only record very vague dates such as "1972" making it difficult to ascertain how long a job actually lasted. A further difficulty involves a person's memory about jobs, and the tendency to expand the amount of time worked. In cases where a past job was verified through the employer, an individual's self report often was not corroborated by company records.

8. Drug Use

There was a discrepancy in twenty-four cases between the data base and the folders regarding drug use history. In seventeen of these cases the data base indicated no drug use and the folder supported a history of drug use. In four cases the data base indicated drug use, while the information in the folders indicated no drug use. Two cases indicated more serious drug use when considering the folder and one case specified the level of drug use

of an unknown drug use as indicated in the data base.

The drug use indicator does not indicate any level of severity. Widely varying cases of drug use were coded the same in the data base. The folders often indicated the intensity of the problem. Severe alcohol, health and counseling needs are also discussed in relation to drug use in the folders while none of this information is included in the data base. Drug use is probably a variable subject to respondent manipulation of information.

VALIDATION STUDY RESULTS
DRUG USE

<u>DATA BASE AND FOLDER COMPARISON</u>		<u>N</u>
Missing and Remains Unknown		2
No Discrepancies		74
Discrepancies		24
<u>Data Base</u>	<u>Folder</u>	
No Drug Use	Drug User	17
Drug User	No Drug Use	4
Drug User	More Serious Drug	2
Unspecified Drug User	Specified Drug Use	1
<hr/>		
TOTAL		100

9. Parole Eligibility Date

There were 15 cases of discrepancies between the data base and folder information regarding a person's parole eligibility date. Since this date is likely to change depending on a person's institutional behavior and the addition of information about the case, the original parole eligibility date is considered, original meaning the date set at the time of commitment.

The data base was just as likely to indicate an earlier parole eligibility date than folder information as a later date. In 8 cases the data base indicated an earlier parole eligibility date and in 7 cases a later parole eligibility date than the offender's folder. In 6 of the 8 cases where the data base indicated an early parole eligibility date, a quick reference index QRI was one of the documents in the folder supporting a later date.

10. Special Education Indicator

Seven individuals had discrepancies between the data base and their folder. In all cases the folder indicated a person had received a GED or attended special classes and the data base did not indicate this. In four cases a person receiving a GED prior to commitment was not noted and in three cases a person attending special classes was not noted in the data base. This information was generally available in probation and classification reports. Only 3 people in the sample were recorded by the data base as attending special classes and 12 people in the sample were recorded as having a GED.

11. State Incarcerations

There were 16 cases where folder information and the data base did not agree on the number of prior state and federal incarcerations. In 10 of these cases the data base indicated fewer prior incarcerations than the folders. Many of these were out-of-state incarcerations. In the other six cases the data base indicated more prior incarcerations than the folder. In 13 of the 16 cases with discrepancies the difference was only one incarceration. The bias toward underreporting prior adult incarcerations could be alleviated by consulting reports

other than Massachusetts probation reports.

12. Juvenile Incarcerations

In only one case was there a discrepancy between the data base and the folder. In that case the data base indicated a single commitment while the folder indicated an original period of incarceration followed by a second period of incarceration after a parole violation.

Since juvenile records may not be included in the folder there is no way of estimating to what extent this variable actually represents a person's juvenile criminal history.

There is an additional problem that a DYS commitment does not necessarily indicate an incarceration. In some cases the place of incarceration may be specified but at other times it will not be.

13. House Incarcerations

In 14 cases there was a discrepancy between the data base and folder information regarding prior house of correction incarcerations. In 9 of those cases the data base reported fewer prior incarcerations than the folder. In 5 cases the data base reported more. As with other incarceration variables out of state sections of a person's criminal career are largely responsible for these discrepancies.

14. Parole Violator Commitment Date

Since the study was done on 1978 commitment file there are very few persons in the file with a parole violator commitment date, 7 in the entire file and 2 in the sample chosen for the study. A person could only have a value for this variable if he was committed

in 1978, released on parole during the year and re-committed as a parole violator to the state system. Because of this sequence of events the number of cases is very small. It should be added that many persons committed the offense that resulted in their 1978 commitment while on parole from a commitment given in another year. This variable, or any other current variable does not permit this to be analyzed. All cases were correctly done on parole violator commitment date as currently defined.

15. Juvenile Paroles

There were 4 cases where there was a discrepancy between the data base and folder information regarding juvenile paroles. While this makes this variable appear to be one of the most valid, most people are coded as having no juvenile paroles and folders often contain no juvenile records, it is not known to what extent a person's juvenile career is actually represented for many individuals.

Of these four cases, in two the data base listed no juvenile paroles and the folder indicated past juvenile paroles (included on Probation and FBI reports). In one case missing data was specified and in the final case the data base listed a juvenile parole that was not supported by folder information. The bias seems to be toward undercounting juvenile paroles as would be expected.

16. Adult Paroles

There were 19 cases with discrepancies between the data base and folder information regarding the number of adult paroles a person had. In 16 of the 19 cases, the data base indicated fewer adult paroles than folder information. In two cases the data base indicated more paroles than the folder. In the final case missing data was available. In most of the cases where the data base indicated fewer paroles the information was not included in the person's probation record. Parole information was represented more fully in classification reports, parole summaries and departmental classification reports. In 14 of the 16 cases where the data base indicated fewer paroles than the folder, the data base indicated no adult paroles.

Summary of Findings

The variables ranged from perfect agreement between the two data sources to forty percent discrepancy. The average number of discrepancies was 13.4 per variable. This appears to indicate a need for increasing the accuracy of many of the variables. When considering discrepancies by case, only 14 cases had no discrepancies. The most discrepancies was one case with nine discrepant variables the average was 1.6 discrepancies per case.

SUMMARY TABLE

NUMBER OF DISCREPANCIES PER VARIABLE

<u>VARIABLE</u>	<u>NUMBER</u>
Time on Job of Longest Duration	40
Last Grade Completed	27
Drug Use	24
Adult Paroles	19
State Incarcerations	16
Date of First Arrest	15
Military Service	15
Parole Eligibility Date	15
House Incarcerations	14
Marital Status	8
Date of Birth	7
Special Education Indicator	7
Juvenile Paroles	4
Race	2
Juvenile Incarcerations	1
Parole Violator Commitment Date	<u>0</u>
TOTAL	214

$$\bar{X}=13.4$$

SUMMARY TABLE

NUMBER OF DISCREPANCIES PER CASE

<u>NUMBER OF DISCREPANCIES</u>	<u>NUMBER OF CASES</u>
0	14
1	25
2	25
3	18
4	12
5	3
6	1
7	1
8	0
9	<u>1</u>
TOTAL	100

Conclusions and Recommendations

Several conclusions can be drawn from the findings of this initial phase of the validation study. First, there is a great deal of discrepancy between the data base and folder information. For all the variables considered there was an average of 13 discrepancies (s.d.=10.7). There was quite a bit of variation among variables. Some had very few or even no cases of discrepancy while others had an inordinately large amount.

Second, there is generally no randomness in the discrepancies, for almost every variable there was a clear bias in these differences. This clear bias has to be of great concern to researchers but can perhaps be solved more readily than random discrepancies.

Finally, as it bears on the issue of whether deviant groups will give information about their behavior, this phase of the study indicates that they do divulge a great deal of information to people within the system. There were very few instances where the most personal information regarding family and criminal background were not given.

Possible courses of action that could improve the validity of this data base are:

1. Missing Data. In many cases discrepancies resulted from missing data being readily available in the folders (time on job of longest duration, military service were particularly affected by this). The problem of missing data can be partially solved by more thorough follow-up of those cases and variables that have missing information, by waiting for appropriate documents to be included in the folders.

2. Use of Additional Documents. In several cases particularly regarding criminal history, use of several documents that may be in the folder would provide more complete information. This is important since there is a bias toward under-counting of prior incarcerations and paroles in the data base.
3. Specifying Variable Definitions. In some cases where discrepancy was limited (e.g. last grade completed) more clear definitions would limit ambiguity and discrepancies.
4. Consistency Checks. In several specific variables as well as to address coding and keypunching accuracy more checks could be introduced. In the case of race using a name check as well as document information would be helpful. Verification of coding and keypunching could be done on a systematic basis.

SOURCES

- Ball, John C., "The Reliability and Validity of Interview Data Obtained from 59 Narcotic Drug Addicts, " American Journal of Sociology, 72: 650-654, 1967.
- Becker, Howard, Outsiders, Free Press, New York, 1963.
- Hindelang, Michael J., Travis Hirschi and Joseph G. Weis, "Correlates of Delinquency: The Illusion of Discrepancy Between Self-Report and Official Measures," American Sociological Review, 44: 995-1014, December, 1979.
- Kinsey, Alfred C., et al., Sexual Behavior in the Human Male, W.B. Saunders Co., Philadelphia, 1948.
- Singer, Robert and Don Gottfredson, Development of Data Base for Parole Decision-Making, Parole Decision-Making Project NCCD Research Center, Davis, California, 1973.
- Williams, J. Allen, "Interviewer Role Performance: A Further Note on Bias in the Information Interview", The Public Opinion Quarterly, 32: 287-294, 1968.