AFWAL-TR-85-4101
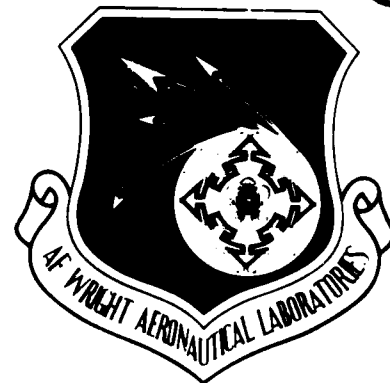
COMPUTER-ASSISTED ANALYSIS OF ELECTRONIC TRANSPORT DATA

Joseph E. Lang, PhD
University of Dayton
300 College Park
Dayton, Ohio 45469

Patrick M. Hemenger, PhD
Electromagnetic Materials Division
Materials Laboratory
Wright-Patterson AFB, Ohio 45433

August 1985

Interim Report for Period January 1982 - June 1985

DTIC FILE COPY

DTIC
SELECTE
JAN 2 1 1987
E

AD-A176 098

87   1 21   097

8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER			
8c ADDRESS (City, State and ZIP Code)		10. SOURCE OF FUNDING NOS.			
		PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
11. TITLE (Include Security Classification) COMPUTER-ASSISTED ANALYSIS OF ELECTRONIC TRANSPORT DATA		61102F	2306	Q1	06

12. PERSONAL AUTHOR(S)
Joseph E. Lang* and Patrick M. Hemenger (*University of Dayton)

13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM 1 Jan 82 TO 1 Jun 85	14. DATE OF REPORT (Yr., Mo., Day) August 1985	15. PAGE COUNT

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB. GR.	semiconductor, silicon, Hall effect, electrical transport, computer data analysis, mobility, effective mass, Hall scattering factor.
20	12		
12	01		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

A non-linear chisquare minimizing computer program has been developed that identifies impurities and dopants in semiconducting materials and calculates their concentrations by fitting temperature dependent Hall data. The program calculates the carrier concentration by solving the charge balance equation and stepping sequentially toward the best fit to the data by varying the concentrations and ionization energies of the impurities. Computer-generated fits to a wide variety of samples have demonstrated the importance of including the correct carrier mass, using the proper Hall scattering factor, employing approximations with discretion, and for some samples including the excited states of the impurity. Emphasis is given to identifying and eliminating sources of error throughout the complete data collection and data analysis sequence.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. P. M. Hemenger	22b TELEPHONE NUMBER (Include Area Code) 513/255-4474	22c OFFICE SYMBOL AFWAL/MLPO

SUMMARY

The electrical behavior of a semiconductor is critical to its usefulness in a broad spectrum of applications ranging from infrared sensitive detectors and microwave amplifiers to high-speed/high-denisty integrated circuits. This behavior is controlled by the presence of dopants, impurities, and defects in the semiconductor and by its temperature. Dopants, which are intentionally added in relatively large concentrations (10^{15} to $10^{18} cm^{-3}$) to create desired properties, must be precisely measured in the same samples as the very low concentrations (less than $10^{12} cm^{-3}$) of unwanted defects and impurities that generally degrade operating characteristics. The only method available for quantitatively measuring the small as well as the large impurity concentrations and determining their identities is based upon the Hall effect. The electrical transport properties, namely resistivity, Hall mobility, and carrier concentration are calculated as functions of temperature by combining resistivity and Hall effect measurements. The temperature dependent carrier concentration data is analyzed to obtain the concentrations and identities of the impurities. Concentrations of less than $5 \times 10^{11} cm^{-3}$ (1 part in 10^{11}) are measured routinely in silicon. The Hall facility used for these measurements has been maintained and employed as a general laboratory analysis tool while it has

been simultaneously upgraded in all respects from data collection through theoretical modeling and analysis.

The precision with which the material parameters can be quantitatively extracted from measurements on the sample is dependent upon the quality of the data and the accuracy of the data analysis procedure. Temperature dependent Hall measurements over a broad range (e.g. 20-400K) are necessary to maximize the information contained in the data, which can make collection of this data very tedious and time-consuming. Automated computer controlled facilities are now routinely employed which do not dramatically reduce the time for a complete data collection run (1/2 to 1 day), but they introduce fewer errors and they do not become bored. This computer control includes automatic setting of the sample temperature with stability to a few millidegrees, adjustment of the magnetic field, setting the sample voltages, and collection of all current and voltage readings. The collected data are used to calculate the carrier concentration as a function of temperature. The identities and concentrations of electrically active centers (dopants, impurities, and defects) are then obtained by using another computer program in an interactive mode which fits the temperature dependent carrier concentration data with a theoretical model by varying the impurity concentations and activation energies as parameters until the difference

iv

between the data and the model is minimized. This fitting process starts with trial values of the parameters introduced into the minimization program by the operator. The ultimate accuracy, and therefore the usefulness, of the results depends upon the precision of the data, the accuracy of the model, the performance of the fitting program, and detailed knowldege of several physical quantities that vary with temperature.

Through a planned and directed thrust over several years to optimize the sensitivity and quantitative accuracy of the Hall analysis technique, all facets of the process have been examined for possible shortcomings and improvements have been incorporated at every stage. This report was written principally to document how information is extracted from data by fitting it with a theoretical model. Successful application of this technique, however, requires a folding together of experiment and theory, therefore this report is also a summary of the Hall analysis work already performed at AFWAL/MLP and an assessment of where further improvements might be made.

Accession For	
NTIS GRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

v

FOREWORD

Because of its scope and time span a number of people made significant contributions to various phases of this project and their participation is sincerely appreciated. Their contributions range from software and hardware development through data collection to theory improvement.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION I

HALL EFFECT MEASUREMENTS

Semiconductors occupy the middle ground between the very good conductors (metals) and the very poor conductors (insulators). This intermediate position, electrically speaking, means that they can be made to take on the characteristics of metals or insulators by altering their environment or preparation. For example, silicon becomes a very good insulator at low temperatures, but can become a conductor at room temperature. The most important property that can be adjusted during preparation to control the conducting behavior of a semiconducting sample is the concentration of specific chemical impurities. Because the electrical and optical properties are intimately related, a semiconductor can also have its optical properties changed. Therefore, in order to accurately describe and control the properties of a semiconductor, the concentration and identities of impurities must be accurately known. For electrically active impurities this analysis is most accurately performed using the Hall effect (References 1 and 2), which is superior to other experimental techniques because it is quantitative as well as very sensitive and accurate.

1.   MEASURED AND DERIVED QUANTITIES

The process of using the Hall effect to determine the impurity concentrations and impurity types in a specific material begins by measuring the resistivity ρ and Hall coefficient R_H over a wide temperature range (see Section I.2). These quantities are related to the carrier concentrations and the carrier mobilities by (Reference 2)

$$\rho^{-1} = e(p\mu_p + n\mu_n) \tag{1}$$

$$R_H = (p\mu_p\mu_{Hp} - n\mu_n\mu_{Hn})/[e(p\mu_p + n\mu_n)^2], \tag{2}$$

where p is the concentration of positive carriers (holes) in the valence band, n is the concentration of negative carriers (electrons) in the conduction band, μ_p and μ_n are the conductivity mobilities for holes and electrons respectively, μ_{Hp} and μ_{Hn} are the Hall mobilities for holes and electrons, and e is the charge on the electron. If conduction in the sample is dominated by one type of carrier, say holes, then these equations reduce to

$$\rho^{-1} = ep\mu_p \tag{3}$$

$$R_H = \mu_{Hp}/(ep\mu_p) = r/(ep), \tag{4}$$

where $r = \mu_{Hp}/\mu_p$ is the Hall scattering factor (commonly called the "r-factor"). The carrier concentration and mobilities may be computed using

$$p = r/(eR_H) \tag{5}$$

$$\mu_{Hp} = R_H/\rho \tag{6}$$

$$\mu_p = \mu_{Hp}/r \tag{7}$$

Unfortunately, only R_H and ρ can be measured directly which means that r must be known in order to calculate the carrier concentration and conductivity mobility precisely. The value of the Hall scattering factor, r, can be measured over only limited ranges in a few cases, and most commonly it must be calculated or estimated. Its value is near 1 and often this has been the estimate used, but recent calculations and measurements (Section V.3) have improved our knowledge of r(T) dramatically.

If measured over a wide temperature range, the carrier concentration data contain information on both the identities of electrically active impurities present in the sample and the concentrations of those impurities. In this report, the

Hall experiment for collecting the carrier concentration data and the computer program for fitting these data will be described. Use of the program to determine the concentrations and chemical species of the impurities present in a sample will be explained. Recent theoretical advances in the calculation of the Hall scattering factor leading to improved determination of the carrier concentration will be discussed. The mobility data can be used to give an indication of the presence of electrically inactive impurities present in the sample and provide information on the crystalline quality of the sample. These data are quantitatively less useful at present because of the greater complication of the theory and the correspondingly greater difficulty in applying it to a concrete situation. Recent theoretical advances have renewed hope that these measurements, too, can be used to characterize samples.

2. THE HALL EXPERIMENT

The Hall experiment involves measuring the voltage at several locations on a sample that is held at a precisely controlled temperature when that sample is subjected to a constant current in the presence and absence of a magnetic field. The sample must be mounted in thermal contact with a fixture which in turn is mounted between the pole pieces of a magnet. The temperature may be varied over the range 4.2 - 400K and the magnetic field must be variable up to at
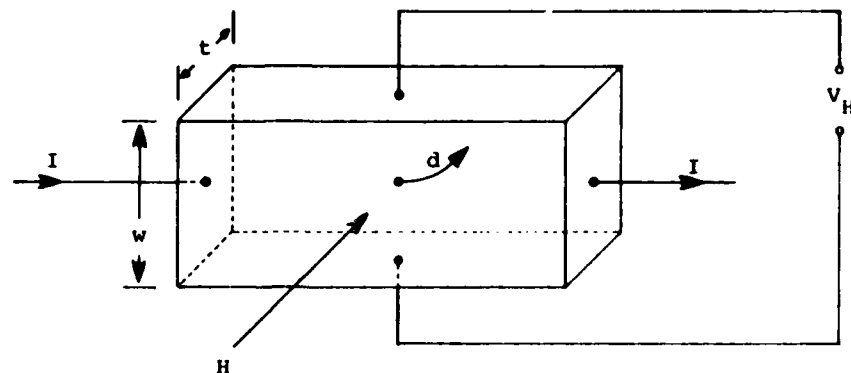
Figure 1. Schematic Hall effect measurement. The letter I
           denotes the current induced in the sample by an
           external power supply, H is the magnetic field
           applied perpendicular to I, V_H is the Hall voltage
           measured perpendicular to I and H, and d is the
           (positive) charge deflected by the Lorentz force
           which is also perpendicular to I and H. The
           sample thickness and width are t and w
           respectively.

least several tenths of a Tesla.

A generalized Hall experiment is illustrated in Figure
1. A current, I, is passed through the sample and the Hall
voltage, V_H, is measured. Initially, the magnetic field is
off, H = 0, in which case V_H = 0. When the magnetic field is
applied, the moving charges are deflected by the Lorentz
force in a direction mutually perpendicular to I and H until
they experience no net force and steady-state is reached.
Deflection of the charges creates a charge concentration

5

gradient across the sample and therefore an electrostatic field. At steady-state, this electrostatic field is called the Hall field and its force on the charges balances the Lorentz force. The Hall voltage, V_H, divided by the distance between the contacts, w, gives the value of the Hall field. The Hall coefficient R_H, is defined as the proportionality constant between the current and Hall voltage and is therefore expressed in terms of directly measured quantities

$$R_H = (V_H/I)(t/H) \tag{8}$$

including the sample thickness, t. Measurement of the resistivity, ρ, is illustrated in Figure 2. The resistivity is the resistance of the material after correction for the dimensions of the sample and can therefore be expressed in directly measurable quantities,

$$\rho = (V/I)(wt/\ell), \tag{9}$$

where V is the voltage measured between contacts a distance apart wher a current, I, is passed through the sample with thickness t and width w.

At low temperatures, the carrier concentration in semiconducting samples decreases due to "freeze-out" which generally results in very high sample resistances and
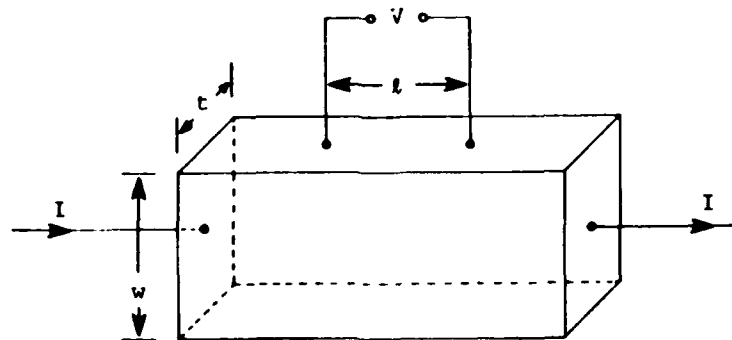
Figure 2.   Schematic resistivity measurement.   The letter I
denotes the current induced in the sample by an
external   power   supply   and   V   is   the   voltage
measured between two contacts spaced a distance
apart.   The sample thickness and width are t and
w, respectively.

therefore small electrical currents. To avoid distorting the

voltages measured at the sample contacts (due principally to

contact resistance), the current drain by the meters must be

much smaller than the current flow through the sample. This

requires meters with very high input impedances that are at

least one hundred times greater than the sample resistance.

To avoid the noise problem introduced by unnecessarily long

RC (resistance-capacitance) time constants, the leads between

the sample and meters must have a cable capacitance that is

operationally equal to zero. This requirement is satisfied

(Reference 3) by connecting electrometers   to the sample

heater coil

beryllium oxide wafer

copper block

sample

magnetic field

heater coil

radiation shield

cold helium gas
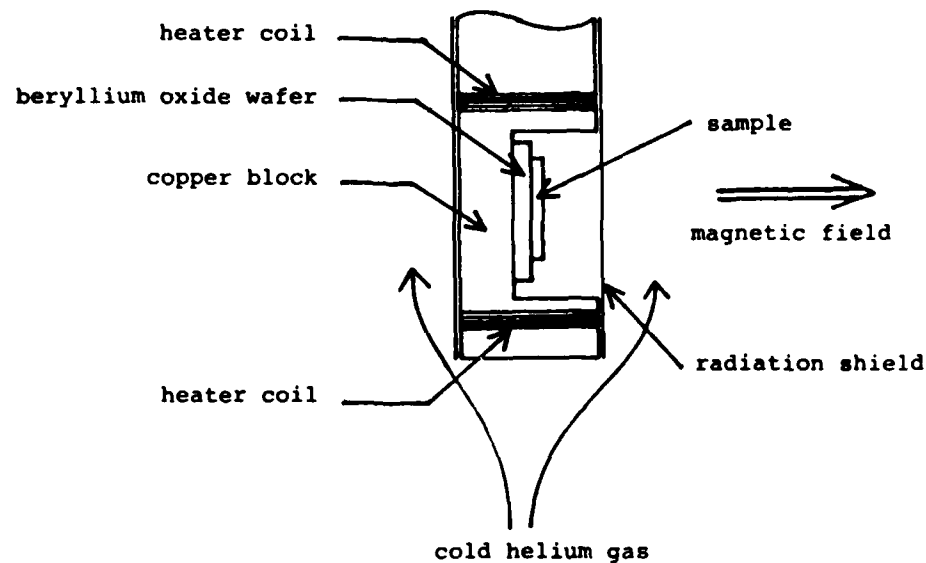
Figure 3. Diagram of the Hall sample holder showing the configuration used for fine temperature control.

voltage probes with triaxial cables and using the unity gain electrometer outputs to drive the inner guard shields to the same potential as the probes. This permits taking reliable voltage readings on samples for which the resistance exceeds 10^{12}ohms. Sample currents at these resistances are on the order of 10^{-11}amps and sensitive electrometers must be used to measure these quantities as well.

For temperature uniformity and control, the sample is held in thermal contact with a beryllium oxide wafer mounted on a copper block as shown in Figure 3. The copper block is in a gently flowing stream of cold helium gas, which acts as a low temperature heat sink. The whole sample region is

covered with a radiation shield and is not in direct contact with the flowing helium. A heater built into the copper can hold the temperature stable to within a few millidegrees. The heater is operated by a closed loop proportional controller which uses either a platinum or carbon-glass resistance thermometer for regulation, but the temperature is actually measured by an independent silicon diode thermometer while the magnetic field is off. The sample holder is part of a liquid helium dewar that is mounted so that the magnetic field is perpendicular to both the sample face and the current flow through the sample. For a given sample, one of two general sample shapes is used: either the Hall bar or the van der Pauw.

a.   Hall Bar

The Hall bar shape most commonly used is that shown in Figure 4. Samples are cut from a wafer using an ultrasonic "cookie cutter". Variations on this shape are used for special cases, but they are all physically equivalent. To collect the resistivity data, a current I is passed through the sample using contacts C_1 and C_2 and a voltage V is measured between contacts R_1 and R_2. From Equation 9 above, the resistivity is given by

$$\rho = (V/I)(wt/\ell), \qquad (9)$$
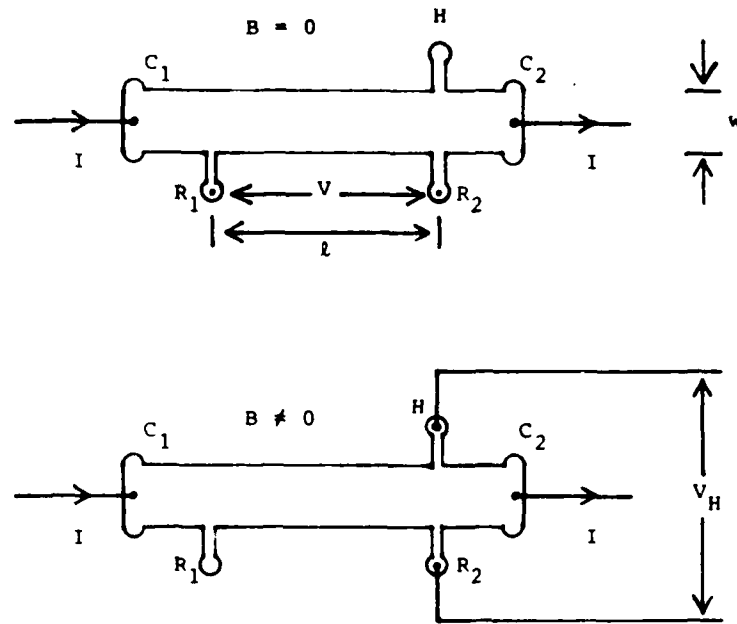
Figure 4. Drawing of a Hall bar showing the connections used to measure the resistivity (top) and Hall coefficient (bottom).

To measure the Hall coefficient, a magnetic field, H, (perpendicular to the page in Figure 4) is applied and a Hall voltage V_H is measured across the sample between contact H and the contact R_2. From Equation 8 above, the Hall coefficient is given by

$$R_H = (V_H/I)(t/H).$$ (8)

Errors due to thermoelectric and thermomagnetic effects as well as to small inhomogeneities and geometrical
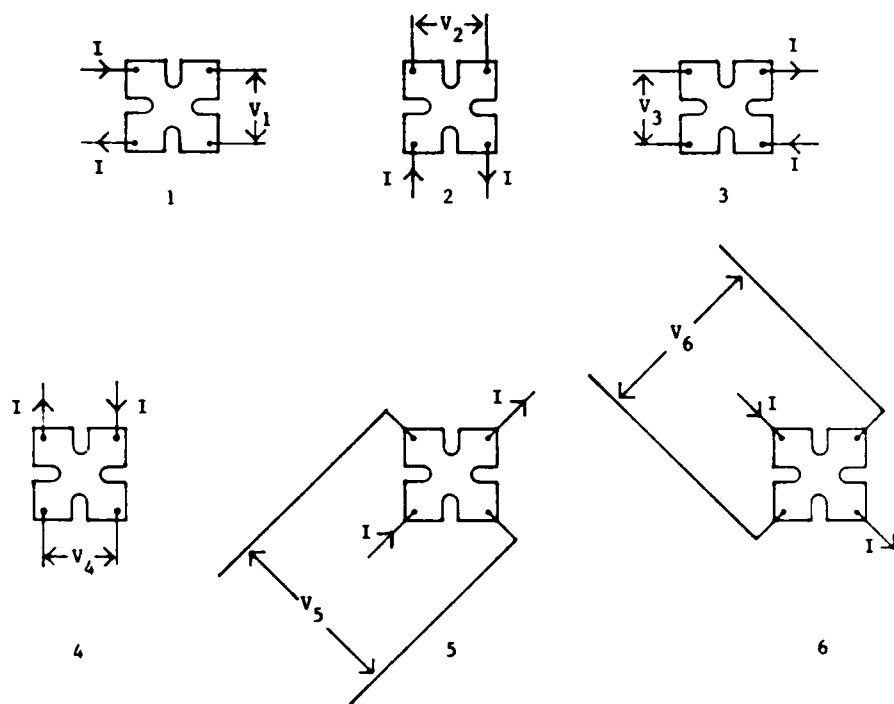
Figure 5.   Illustration of the Six Possible Configurations of
Electrical   Connections   Used   to   Measure   the
Resitivity (1,2,3,4) and Hall Coefficient (5,6) of
a van der Pauw Sample.

irregularities   are   minimized   by   collecting   data   for   all

combinations of current and magnetic field reversals and then

averaging suitably.

b.   Van der Pauw

Like   the   Hall   bar,   the   general   van   der   Pauw

(References 4 and 5) shape must be homogeneous and have

uniform thickness, but it also must contain no holes and have

four point-size contacts on the periphery.   Data collection

and analysis are much simplified if a symmetrical shape is

used and the contact size restrictions are greatly relaxed if some form of the "clover leaf" is chosen for the sample shape, Figure 5. The slots in the clover leaf effectively restrict the sample's measured area to the central portion -- a small area about 2mm square -- and electrically place the contacts far from this active region. The advantage is twofold: first, the contacts can be made much larger without introducing error (Reference 6) and second, a small sample area permits sampling of wafers with higher spatial resolution.

Data collection using a van der Pauw sample differs from the Hall bar primarily in the fact that each contact plays the role of both voltage probe and current connection during the necessary switching process that is shown schematically in Figure 5. Measurements are required at only three of the six configurations, say 1, 2, and 5 or 3, 4, and 6. For the 1, 2, 5 set, one puts the current I through the sample and measures the voltage V_1 in position 1, V_2 in position 2 and then V_5 with the magnetic field on in position 5. To reduce these measurements to resistivity and Hall constant, one must assume that V_5 is zero when H is zero. This is rarely the case, and so, in analogy with the Hall bar case above, V_1, V_2, and V_5 are remeasured and suitably averaged with the values obtained when the current and/or the magnetic field is reversed. Therefore, a complete set of data at each

12

temperature consists of V_1, V_2, V_5, I and H. For more complete averaging, equivalent data can be taken in the other positions and would include V_3, V_4, V_6, I and H.

The resistivity may be calculated (References 4 and 5) using

$$\rho = \pi t (R_1 + R_2) f(R_1/R_2)/(2\ln2),$$
(10)

where t is the thickness of the sample and R_1 and R_2 are given by

$$R_1 = V_1/I$$
$$R_2 = V_2/I.$$
(11)

The function $f(R_1/R_2)$ is a dimensionless quantity dependent only upon the ratio of the resistances measured for configurations 1 and 2 and equals one when R_1 equals R_2. This is the case for an electrically isotropic geometrically symmetric sample. The ratio R_1/R_2 is monitored as a function of temperature and it should remain constant. If it does not, material inhomogeneity is implied.

The Hall constant is found using

$$R_H = (\Delta V_5/I)(t/B),$$
(12)

13

where

$$\Delta V_5 = V_5(B) - V_5(B=0). \tag{13}$$

14

SECTION II

THE CHARGE BALANCE EQUATION

The carrier concentration in a semiconducting sample is determined (References 2 and 7) by assuming charge neutrality in the sample, a condition expressable in terms of the charge balance equation:

$$p + \sum_{\ell} N_{D\ell}^{+} = n + \sum_{i} N_{Ai}^{-} \tag{14}$$

Here, p denotes the concentration of holes (positive carriers) in the valence band of the sample, n the concentration of electrons (negative carriers) in the conduction band of the sample, $N_{D\ell}^{+}$ the concentration of positively charged (ionized) donors (typically group V elements in silicon) in the sample, and N_{Ai}^{-} the concentration of negatively charged (ionized) acceptors (typically group III in silicon) in the sample. The sums are over the different species of donors and acceptors.

For a silicon sample more heavily doped with acceptors than donors, the number of free electrons in the sample is negligible at and below room temperature and all of the donors in the sample end up being positively charged because their outer electrons have moved away to acceptor sites, compensating them. The sum over donors may be replaced by a

15

single number N_D, which represents the total contribution of all of the donors in Equation 14. Then, the electrical conductivity is dominated by the holes and the sample is said to be p-type. For this case, the charge balance equation takes the simpler form:

$$p + N_D = \sum_i N_{Ai}^- \qquad (15)$$

If Fermi statistics are applied to the acceptors, and Boltzmann statistics are applied to the concentration of holes in the valence band (this approximation is discussed in Section V.1.b), Equation 15 takes the form:

$$p + N_D = \sum_i N_{Ai}/(1 + p/\Phi_i) \qquad (16)$$

where

$$\Phi_i = N_v/(\sum_j g_{ij} \exp(E_{Aij}/kT)) \qquad (17)$$

Here, the sum over j includes the ground state and the excited states of the impurity. The degeneracy of the state is given by g_{ij} and the (positive) binding energy of the state is given by E_{Aij} and is measured from the valence band edge toward the conduction band. The density of states in the valence band is denoted by N_v.

16

The carrier concentration $p(T)$ is measured as detailed above, and the parameters $N_v(T)$, g_{ij} for all of the states, and E_{Aij} for the excited states are found from theoretical analysis and other experiments. By fitting the solution of Equation 16 to the measured $p(T)$ data, the value of N_D, and the values of N_{Ai} and E_{Ai0} (ground state ionization energy, denoted below by E_{Ai}) may be determined. That is, the concentration of donors, the concentration of each acceptor, and the ionization energy of each acceptor are found. The ionization energy of an acceptor serves to identify it. It should be noted that unless the sample is heavily doped or the ionization energy E_{Ai} is very small, the excited states contribute very little to the sum in Equation 17 and they may therefore be safely neglected. Then, the temperature dependent function Φ_i takes the form that is applicable in most cases:

$$\Phi_i = N_v \exp(-E_{Ai}/kT)/g_{i0}. \qquad (18)$$

The degeneracy of the ground state, g_{i0} can be treated as an unknown parameter during the fitting process so that it may be determined from the $p(T)$ data. When this test gives non-integral values for g_{i0}, it may indicate that the excited states are contributing significantly in Equation 17.

17
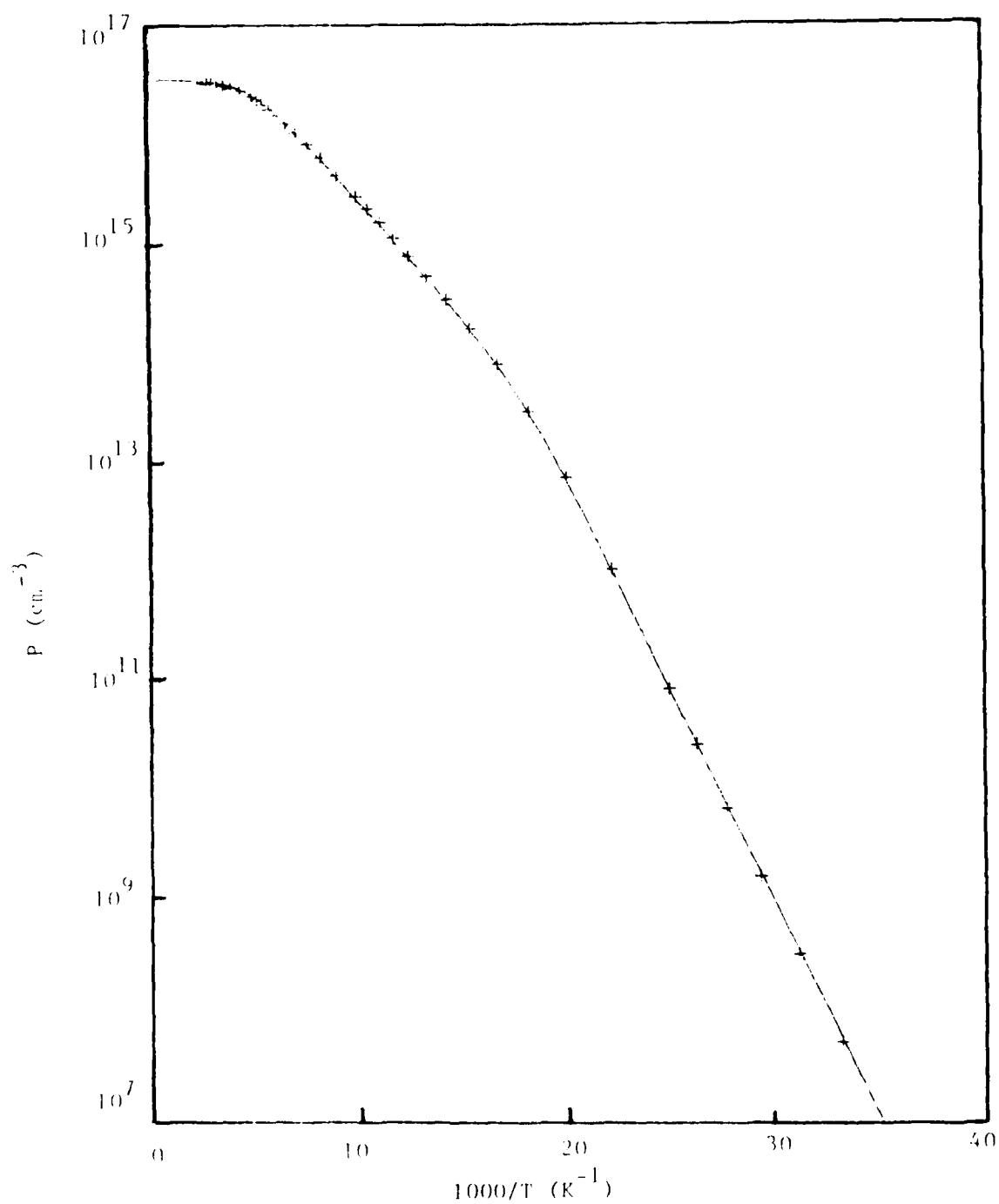
Figure 6. Concentration of holes versus 1000/T for a sample with one undercompensated acceptor species. The curve fitted to the data points (+ signs) gives a reduced chisquare of 1.3.

18

1. SINGLE ACCEPTOR SOLUTION

The graph in Figure 6 shows the typical form of the hole concentration data (Reference 8) for a sample with only one type of acceptor. Note the three distinct portions of the curve. At the highest temperatures on the left, the hole concentration is a constant independent of temperature. This "exhaustion region" develops when the temperature is high enough to cause thermal ionization of all acceptors in the sample, then, N_A^- in Equation 15 is just N_A and the hole concentration is independent of temperature:

$$p = N_A - N_D. \tag{19}$$

At the lowest temperatures on the right, in the "freeze out region," the carriers become bound to the acceptors and the hole concentration is given by:

$$p = (N_A - N_D)\Phi/N_D \propto \exp(-E_A/kT). \tag{20}$$

This produces the long linear section of the curve at low temperatures with a slope related directly to the ionization energy of the acceptor. Between these two temperature extremes, there may be a linear section of the curve with a slope of half the low temperature slope where the hole concentration is

$$p = \sqrt{N_A \Phi} \propto exp(-E_A/2kT) \qquad (21)$$

and the extent of this intermediate region depends upon the ratio N_A/N_D. The half-slope region becomes very long for large values of the ratio N_A/N_D and the break between this region and the low temperature region occurs for a carrier concentration near the value N_D.

Equations 19, 20, and 21 may be used as a guide when examining graphs of carrier concentration data. The logarithm of the carrier concentration is plotted versus 1000/T to straighten the graph in the intermediate and low temperature regions whe.e the data are exponential. The presence of a clearly visible intermediate temperature region in Figure 6 indicates that the concentration of donors is at least an order of magnitude less than the concentration of acceptors in this sample and the position of the break to the low temperature slope gives an estimate for N_D. The height of the curve in the exhaustion region gives an approximate value of $3.0 \times 10^{16}/cm^3$ for the concentration of acceptors, and the slope of the low temperature data give an ionization energy of 0.076 eV which is the approximate value for gallium. This quick scan of the data gives important information, namely that gallium is most probably the dominant impurity at a concentration of about $3.0 \times 10^{16}/cm^3$,

and that the net concentration of donors is roughly $3 \times 10^{13}/cm^3$. A detailed computer analysis of this sample will be presented in Section III.

When more than one type of acceptor impurity is present in the sample, the curves take the same general form, and while the slopes of the sections may be either proportional to the ionization energy of the acceptor or half that value, the interpretation of the curve is much more difficult. In this case, a nonlinear fitting program is needed to disentangle the information from the various portions of the curve. Adding to the complication is the fact that N_D and the N_{Ai} include only impurities which contribute to the measured $p(T)$. "Invisible" impurities may be present which form defects that are not electrically active as well as three categories of donors and acceptors that are not detectable: (1) deep acceptors which have an ionization energy E_{Ai0} that is greater than the highest experimental kT and are therefore not ionized, (2) shallow acceptors with ionization energies less than the energy of the lowest partially compensated acceptor that are overcompensated by a factor of 5 or more, and (3) the portion of N_D required to exactly compensate the invisible shallow acceptors in category (2). The measured portion of N_D is referred to as the net donor concentration. In other words, when a complete Hall analysis is performed, it is possible that the sample

21

may contain some acceptor centers, both shallow and deep, and some donor centers in addition to those detected. This is the case for silicon samples intentionally doped with indium which always contains some boron. If any of the boron is electrically active (i.e. undercompensated) extra donors may be introduced intentionally to overcompensate the boron and therefore make it electrically inactive in order to reduce the low temperature carrier concentration.

2. MULTIPLE ACCEPTOR SOLUTION

Data from a sample with more than one type of electrically active acceptor (Reference 9) are shown in Figure 7. The more complicated shape of this curve indicates that there are multiple acceptors in the sample. A simple fit to these data indicates the presence of two acceptors whose energies are determined by the straight line portions of the data. The steepest portion of the curve at high temperature has a slope indicative of the presence of indium in the sample, and the less steeply sloped portion at low temperature indicates the presence of aluminum. The detailed shape of the curve, however, indicates the presence of a third acceptor with an ionization energy between that of indium and aluminum. This acceptor is known as indium-X and including it in the analysis improves the fit to the data.

It is a very general property of carrier concentration
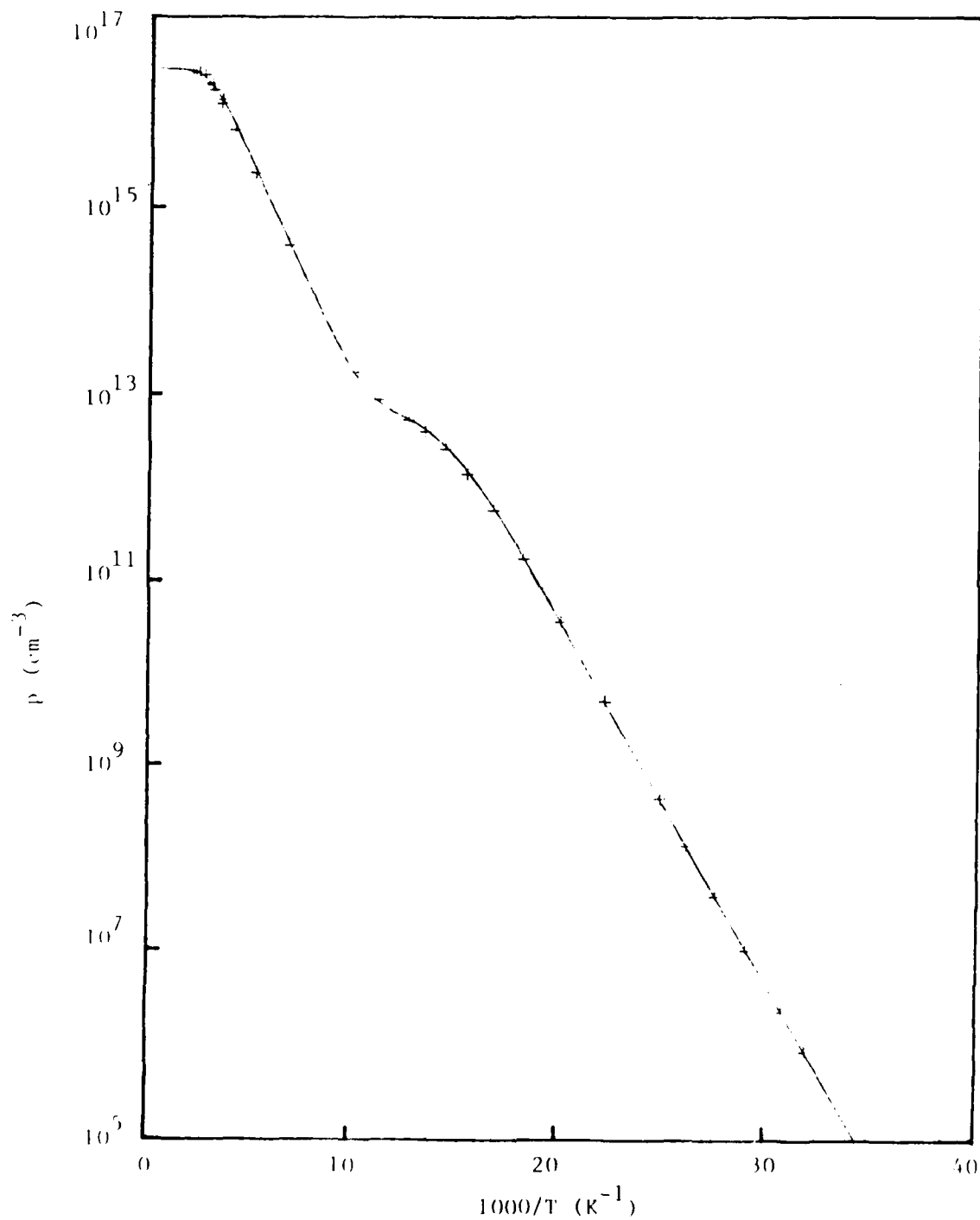
Figure 7. Concentration of holes versus 1000/T for a sample with several
undercompensated acceptor species. The two-acceptor fit shown has
a reduced chisquare of 3.4 whereas a three-acceptor fit has a
reduced chisquare of 0.9.

23
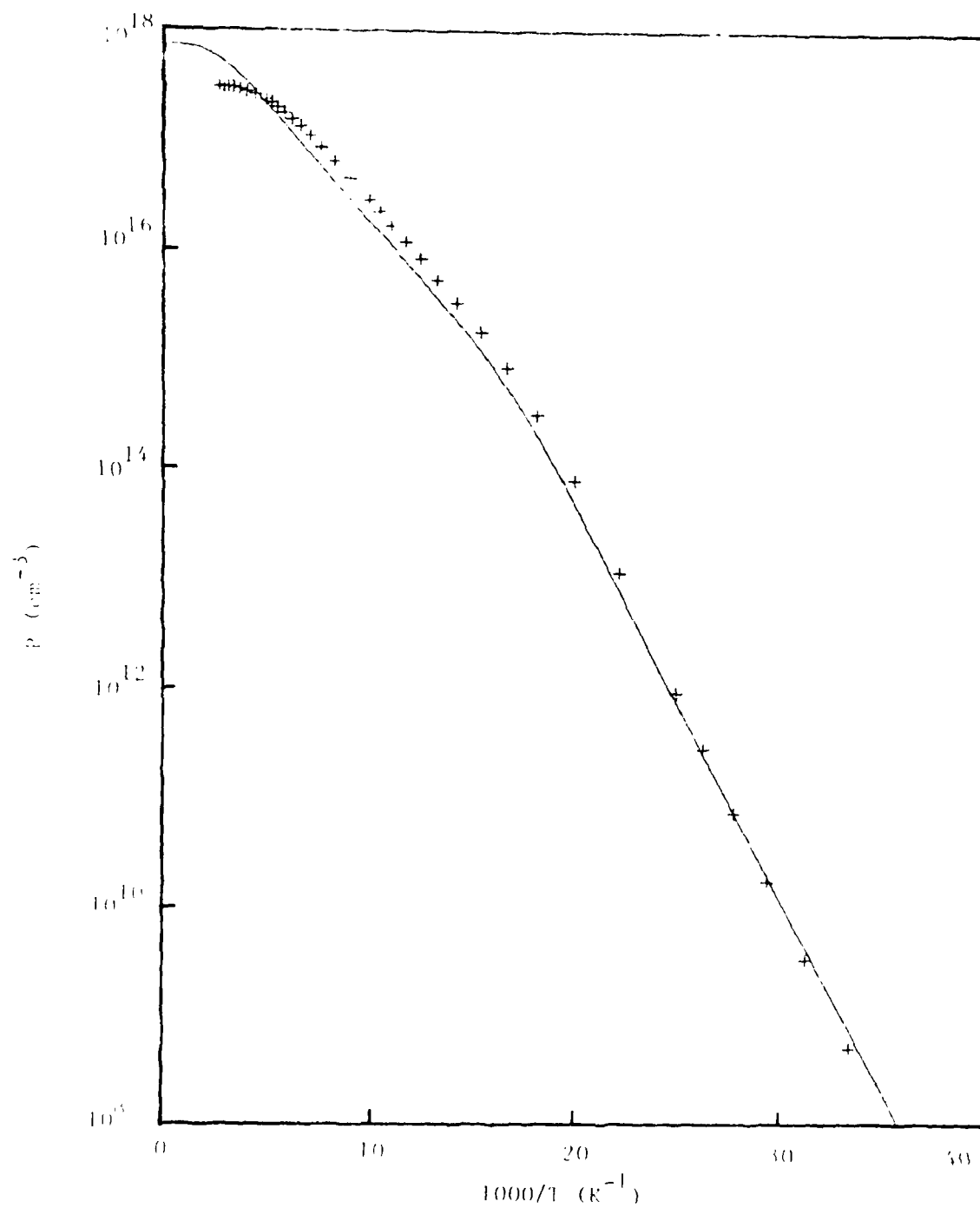
Figure 8. Data of Figure 6 after multiplication by a factor of 10. The fitted curve shown has a reduced chisquare of 128.2.

24

data that the detailed shape of the curve contains information beyond that available in just the slopes of the straight line sections and the absolute height of the curve. This is clearly seen by comparing the one acceptor fit to the data in Figure 6 with a one acceptor fit to the same data multiplied by a factor of 10 as shown in Figure 8. The higher values of the data require higher values of the impurity concentrations, but even so, these higher values cannot reproduce the detailed shape of the curve: the shape retains the information that the true impurity concentrations are an order of magnitude lower. Therefore, even in the simple one level situation, the maximum amount of information can only be extracted from carrier concentration data by performing a computer fit to them. However, if a particular data set contains only a single slope with no bends or structure, then the data can be shifted in magnitude with no adverse effect upon the fit and the only information contained in the data is one energy level value.

For the case of two acceptor species, the exhaustion region again develops at high temperatures where all of the acceptors of both types are ionized. Then, Equation 15 has the solution

$$p = N_{A1} + N_{A2} - N_D. \qquad (22)$$

For temperatures immediately below the exhaustion region, the solution of the charge balance equation takes the form

$$p = \sqrt{N_{A1}\Phi_1} , \qquad (23)$$

where the acceptor listed is the one with the deeper ground state (greatest absolute value of the ionization energy), and the slope in this region is $E_{A1}/2$. If N_D is greater than N_{A2}, the shallower acceptor is overcompensated, and the carrier concentration in the low temperature regime is

$$p = (N_{A1} + N_{A2} - N_D)\Phi_1/(N_D - N_{A2}) \qquad (24)$$

giving a slope of E_{A1}. This looks exactly like the one accceptor solution with an effective number of donors given by $N_D - N_{A2}$. This important result, namely that overcompensated shallow acceptors are not "seen" in the data was remarked upon above. Note that the concentration of donors found in the analysis is actually the concentration of donors minus the concentration of the overcompensated shallow acceptors.

If the shallower acceptor is exactly compensated, i.e. $N_D = N_{A2}$, the solution for high and intermediate temperatures stays the same as above, but the solution at low temperature takes the form

26
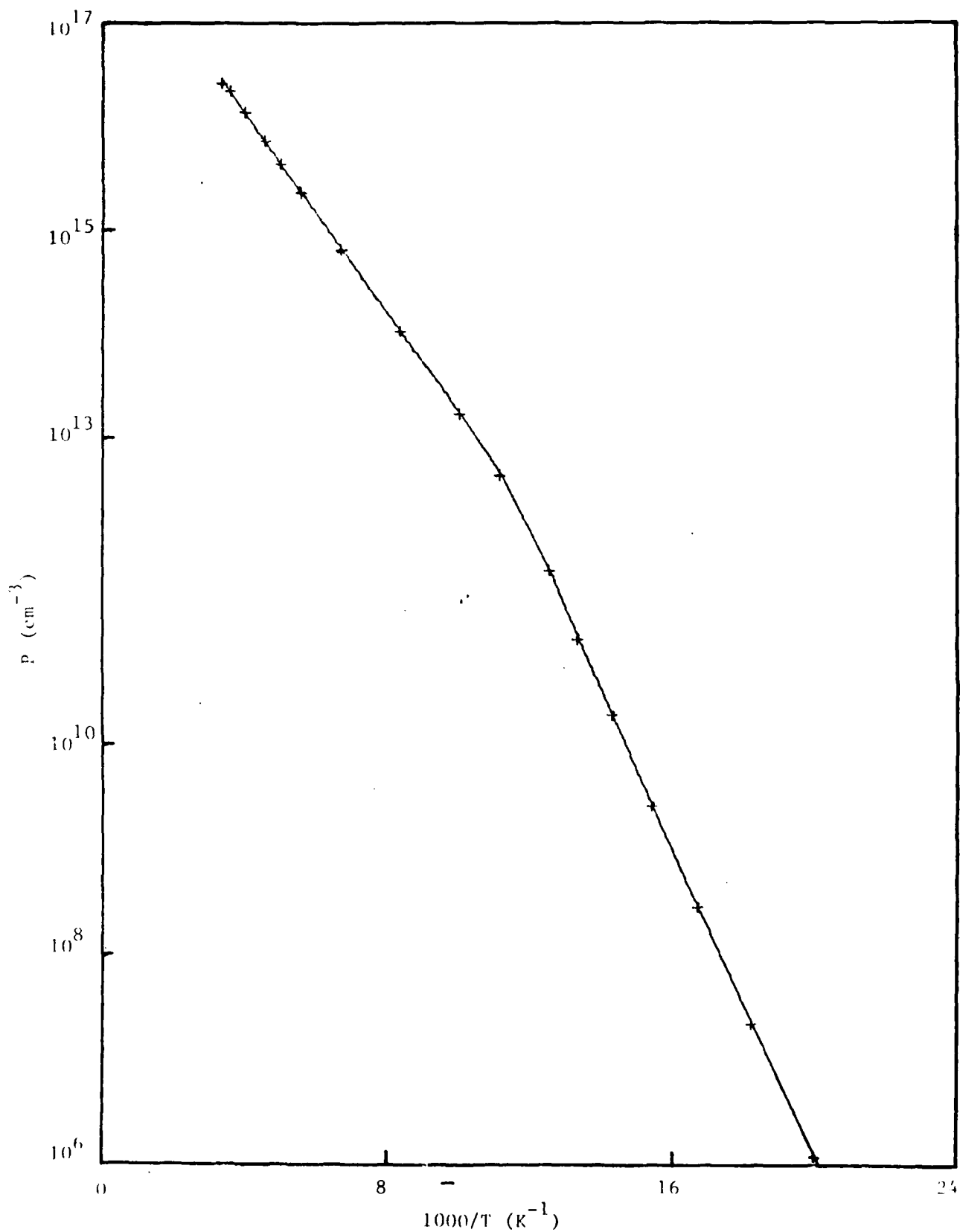
Figure 9. Hole concentration versus 1000/T for a sample doped with two acceptor species. For the fit shown a reduced chisquare of 1.8 was obtained assuming $r = 1$, a net donor concentration of $4.4 \times 10^{12}/cm^3$ and a concentration of $5.7 \times 10^{14}/cm^3$ for the shallow acceptor.

27
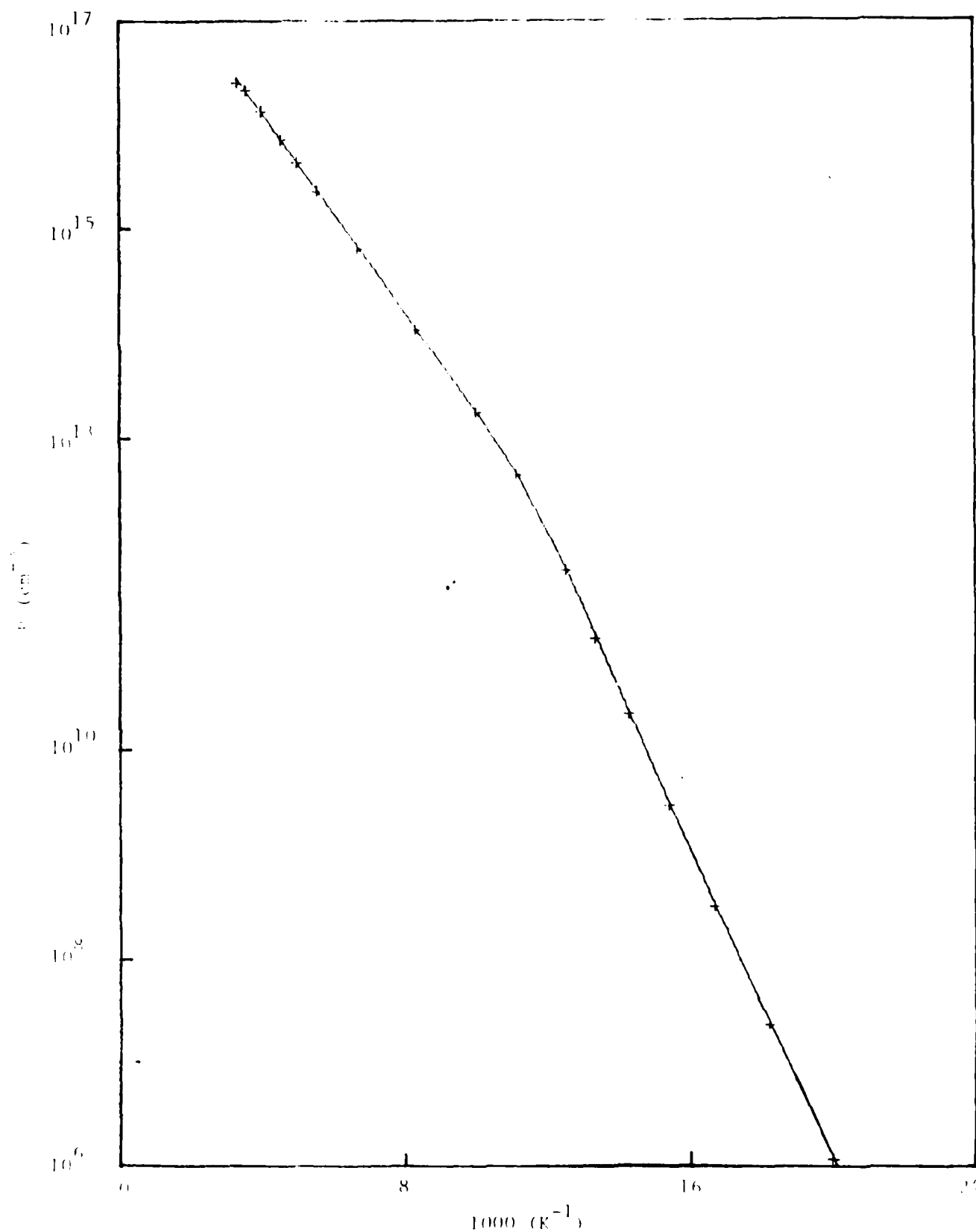
Figure 10. Fit to the data of Figure 9 using a closely compensated shallow acceptor. For this fit a reduced chisquare of 1.7 was obtained assuming $r = 1$, a net donor concentration of $3.8 \times 10^{12}/cm^3$ and a concentration of $3.9 \times 10^{12}/cm^3$ for the shallow acceptor.

$$p = \sqrt{N_{A1} \Phi_1 \Phi_2 / N_D} \ , \tag{25}$$

and the low temperature slope is $(E_{A1} + E_{A2})/2$. If the shallower acceptor is undercompensated, i.e. N_D less than N_{A2}, the solution at high and intermediate temperatures is again the same as above, but the solution at low temperatures is

$$p = (N_{A2} - N_D) \Phi_2 / N_D \tag{26}$$

giving a low temperature slope of E_{A2}.

3. ALIASING

If the value $(E_{A1} + E_{A2})/2$ is near the ionization energy of another impurity, say E_{A3}, two different fits to the data will be possible, one with the impurities #1 and #3 showing up in the fit, and one with impurities #1 and #2 showing up, but where the concentration of donors is almost exactly equal to the concentration of acceptor impurity #2. This possibility of the aliasing of one impurity by another is shown in Figures 9 and 10. Both show the same data (Reference 10), but the first shows a fit with a shallow acceptor that is undercompensated, while the second shows a fit with a shallow acceptor that is almost exactly compensated. Note that in both of these cases, the presence

29

and character of the deeper acceptor is unambiguous. The only question to be determined is whether the shallower acceptor is closely compensated or not. Larrabee (Reference 11) performed a series of experiments on indium doped silicon to determine whether the indium-X acceptor was the result of close compensation of a shallow level. By varying the donor concentration, he was able to show that close compensation of a shallow level could not explain the data. It is a curious fact that several acceptors have ionization energies near the average of the binding energies of two other acceptors (Reference 12). It is especially important to perform a careful analysis such as that done by Larrabee for these levels (References 13 and 14).

SECTION III

DETERMINATION OF FIT QUALITY

The quality of a computer fit to experimental data can be judged in a number of ways. The most important of these is a quantitative statistical measure called the reduced chisquare which depends on the mismatch between the values of the fitting function and the data points. Though it is a quantitative indicator, it gives only aggregate information about the fit as a whole and does not provide detailed information about the relationship between the shape of the curve and the trend of the data. Detailed information of this type may be obtained by examining a graph containing the data points and the curve produced by the fitting function to see whether the curve faithfully reproduces the bends and points of inflection of the data. It may be necessary as well to examine numerical values of the data and fit to look for systematic or correlated errors which are too small to see on a graph. When the fit seems to be a good one according to these methods, it may still be producing physically unreasonable values of the fitted parameters, a sign that the fit is a poor one. Even if the values of the parameters are reasonable, one or more of the parameters may be determined only approximately by the fit. In this case, the data is not constraining the parameter(s) strongly and the fit is said to be "mushy", meaning that the value of

31

chisquare is only weakly dependent on the value of the
parameter(s) in question.   In this section, each of these
measures will be examined in turn to show the information
they give concerning the quality of a fit and the action
which may be taken to improve the fit, if needed.

1.   REDUCED CHISQUARE

The fitting of a theoretical formula with adjustable
parameters to a set of data is done by varying the parameters
so as to minimize a quantity that measures the distance of
the fitting curve from the data points.   The appropriate
measure is called chisquare and is defined by (Reference 15)

$$\chi^2 = \sum_{i=1}^{N} [y_i - f(x_i)]^2 / \sigma_i^2 \qquad (27)$$

where N is the number of data points, y_i is the ith data
value with standard deviation σ_i, and $f(x_i)$ is the value of
the fitting function for the ith data value.   If a fitting
function could be found that went through all of the data
points exactly, the value of chisquare would be zero.

Ordinarily however, the data are randomly displaced
about the curve due to the finite precision of the experiment
and the minimum value of chisquare is not zero.   A fit using
a function $f(x_i)$ is judged to be a poor one if it produces a
"large" minimum chisquare and is judged to be good if it

32

produces a "small" minimum chisquare. If there are n free parameters in the fitting function, we might expect that the fitting function could go through n of the data points exactly and miss the rest by about one standard deviation. As an estimate, then, the minimum value of chisquare would be approximately

$$\chi^2 \sim N - n \tag{28}$$

Because this minimum value depends only on the number of data points and the number of free parameters, the significant quantity is not chisquare, but chisquare divided by N - n,

$$\chi_\nu^2 = \chi^2/(N - n) \tag{29}$$

called the reduced chisquare. Its minimum value should be about one, and therefore, a poor fit is one with a reduced chisquare much larger than one.

In the fits shown in this report, the standard deviation, σ_i, is assumed to be 3% of the data value:

$$\sigma_i = 0.03y_i. \tag{30}$$

If the standard deviation is larger than that given by Equation 30, or if the fitting function is inappropriate for

33

the data, the smallest value of the reduced chisquare will be greater than one. On the other hand, if Equation 30 overestimates the deviation of the data and if the theory describes the data well, the reduced chisquare will be less than one. When the minimum reduced chisquare is too large for a given fit, parameters may be added to the fitting function to improve the fit. For the hole concentration data discussed here, this may be accomplished by adding more acceptor levels to the fitting formula. The data in Figure 7, for example, are better fit assuming that there are three rather than two acceptor species doping the sample.

One might think that adding extra parameters in the fitting function could always improve the fit and decrease the minimum value of the reduced chisquare. While the minimum value of chisquare itself will decrease as a result of adding extra parameters to the fitting function, the minimum value of the reduced chisquare will not get smaller because the value of the denominator in Equation 29 also decreases as n is increased. As parameters are added during a fitting session, the minimum reduced chisquare will first decrease and then increase, having a lowest value as a function of n. For this lowest value, the maximum information is extracted from the data. Practically speaking, parameters are added so long as the results are physically reasonable.

34

TABLE 1

RESULTS FOR THE DATA IN FIGURES 4 AND 6

	Net Donor Concentration (cm^{-3})	Acceptor Concentration (cm^{-3})	Acceptor Energy (eV)	Reduced Chisquare
Fig. 6	7.17×10^{13}	3.17×10^{16}	0.071	1.3
Fig. 8	8.46×10^{14}	7.16×10^{17}	0.066	128.2

Compare the values in Table 1 for the reduced chisquare for Figures 6 and 8. The fact that the value for Figure 6 is near 1.0 is an encouraging indication that this is a good fit to good data, whereas the large value for the same type of fit to the 10X data in Figure 8 leads to a rejection of this fit without giving information as to the cause of the problem.

The use of the reduced chisquare to measure the quality of two different fits to the same data is illustrated in Figures 11 and 12 with (respectively) a two acceptor and a three acceptor fit to the data (Reference 16). The unusually low value of the reduced chisquare for the three acceptor fit, 0.15, is an indication of the excellence of the data. The fact that the reduced chisquare for the three acceptor fit is less than the reduced chisquare for the two acceptor fit, 3.65, lends statistical support to the hypothesis that there are actually three different acceptors in the sample. This reduction in chisquare does not always happen, for as

35
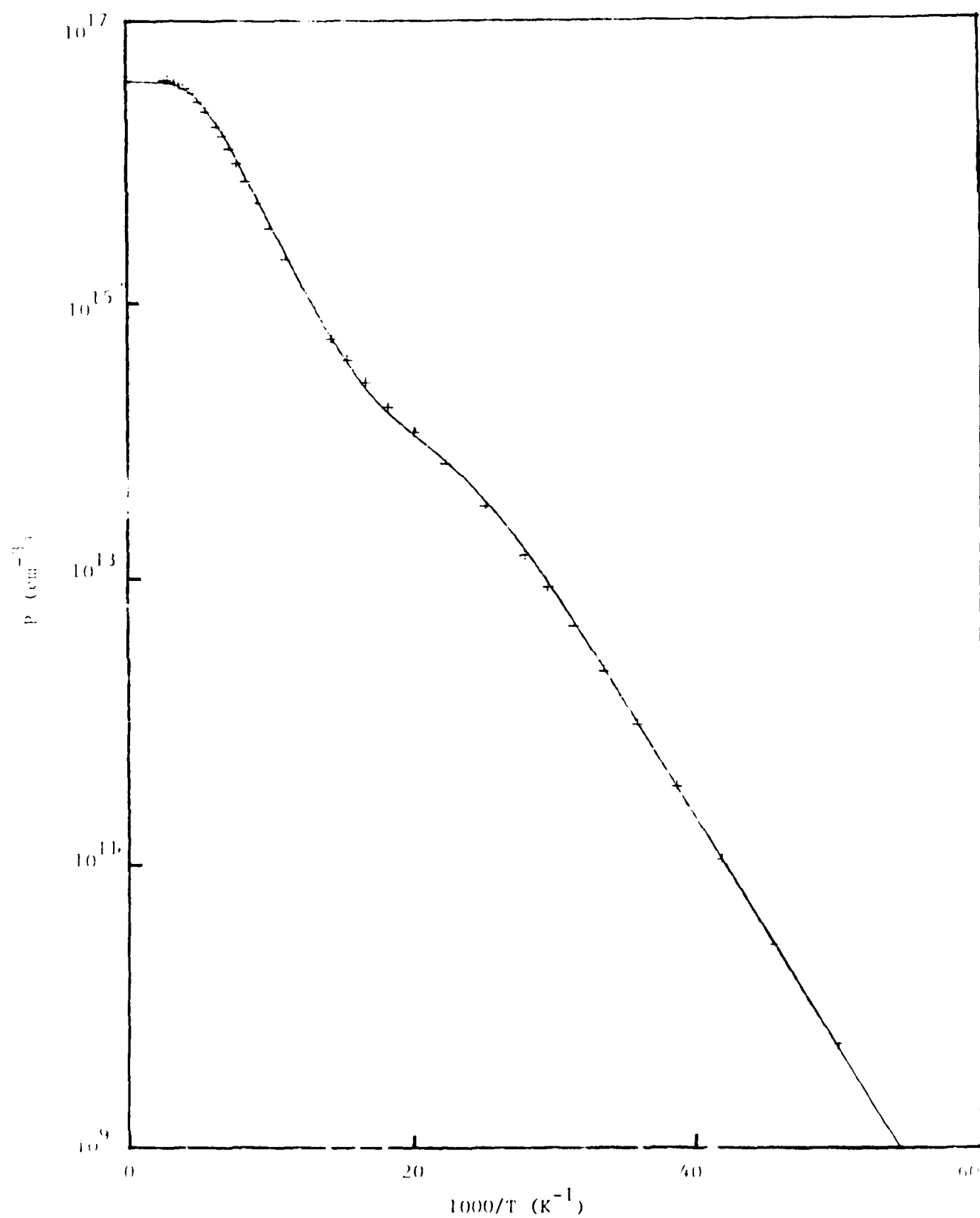
Figure 11. Hole concentration versus 1000/T for a sample doped with three
acceptor species. The two acceptor fit shown has a reduced
chisquare of 3.65.

36
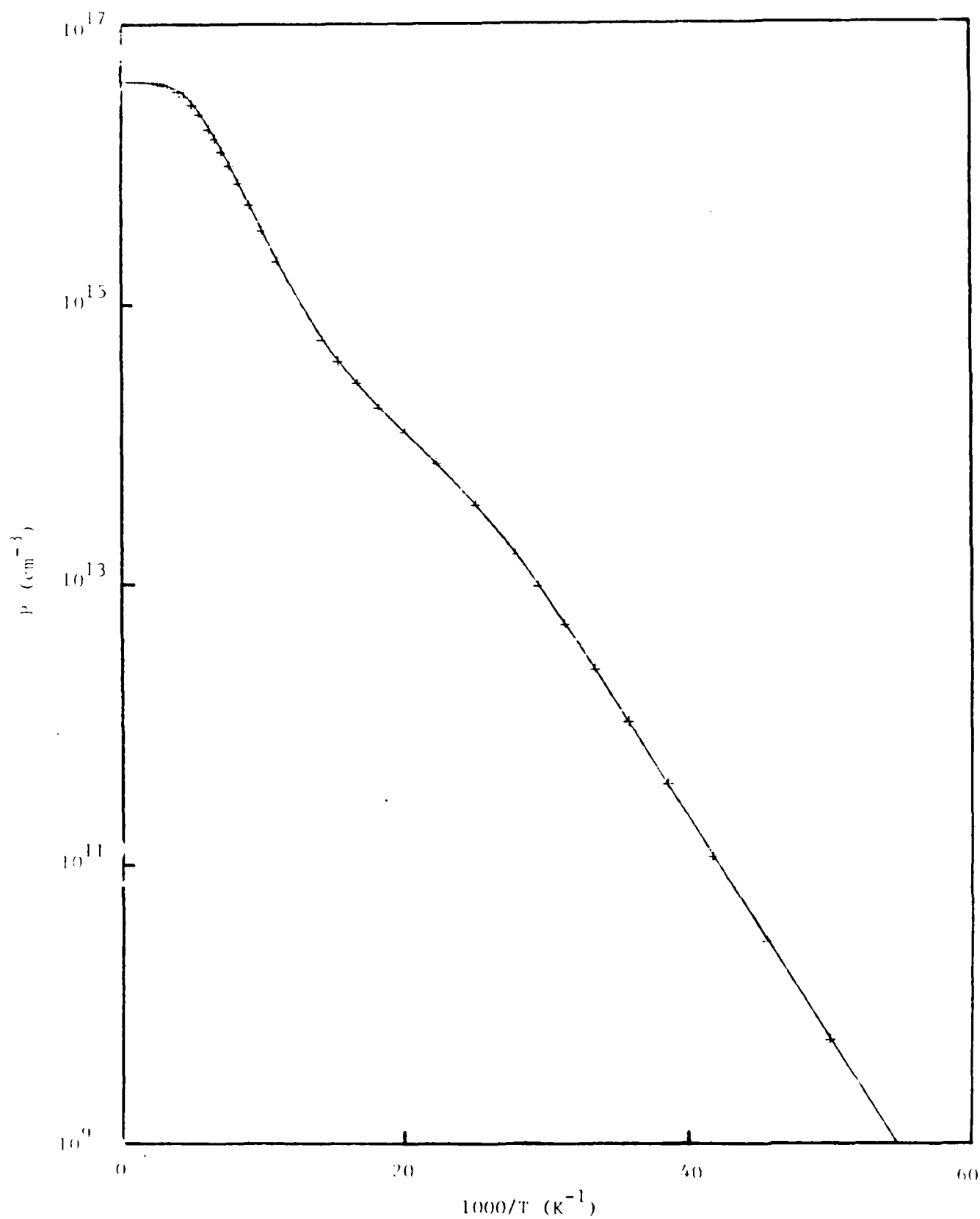
Figure 12. Three acceptor fit to the data of Figure 11 having a reduced chisquare of 0.15.

discussed above, increasing the number of parameters by increasing the number of acceptors in the fitting process may not lead to a smaller reduced chisquare.

2.   GRAPHICAL COMPARISON AND ERROR ANALYSIS

The value of the reduced chisquare does not tell the whole story about the quality of a fit. If a perfect theory is used to fit a set of data that has only random error in it, the minimum value of the reduced chisquare would reflect only the residual random error in the data. That is, the data points would be displaced above and below the curve randomly. In a realistic situation, however, there will be systematic error in the data and perhaps the theory. Then the minimum value of the reduced chisquare has two components: a part due to the random error in the data which shows up as described above, and a part due to the systematic error which shows up as a correlated displacement of the data points above or below the curve.

For example, the fit shown in Figure 8 is known to be poor because of the high minimum value of the reduced chisquare. The poorness of the fit may be confirmed by comparing the fitted curve to the data points over a range of temperatures. The difference between the data and the fit is systematically either high or low over a temperature range, the largest percent difference being about 30%, and this is

38

evidence that the poor fit is caused by systematic error. In this particular case, the poor fit was caused artificially by multiplying the accurate data of Figure 6 by a factor of 10 to produce the data in Figure 8. In other words, a correlation of the error from point to point is an indication of systematic error, the amount of systematic error being far larger than the irreducable systematic percent difference between the data and the best fit.

Systematic error in an experiment can come from a wide variety of sources. Experience leads to the identification of the temperature, magnetic field strength and sample thickness as the greatest experimental sources of systematic error in Hall effect experiments. If the theory is approximate, it will contain systematic errors and the fitted values will differ from the experimental values for this reason. For example, as noted above, the usual estimate for the value of the Hall scattering factor r is to take $r = 1$. It is known that r can differ from this value by as much as 30%. In Table 2, the fitted values are compared to the experimental values of the carrier concentration for the data in Figure 6. Note that even for these data there are small but systematic deviations of the fitted values from the experimental values over wide temperature ranges. This leads to the conclusion that either the theory or experiment (or both) has residual systematic error in it. It can be very

39

TABLE 2

COMPARISON OF DATA AND FIT FOR FIGURE 6

T (K)	Carrier Concentration Data (cm^{-3})	Carrier Concentration Curve (cm^{-3})	Percent Difference
30	5.10×10^7	5.36×10^7	2.5
32	3.26×10^8	3.46×10^8	3.0
36	7.05×10^9	7.20×10^9	1.0
40	8.81×10^{10}	8.59×10^{10}	-1.3
45	1.07×10^{12}	1.01×10^{12}	-2.9
55	2.94×10^{13}	2.89×10^{13}	-0.8
65	1.69×10^{14}	1.66×10^{14}	-0.9
80	7.76×10^{14}	7.53×10^{14}	-1.5
100	2.73×10^{15}	2.62×10^{15}	-2.0
130	8.19×10^{15}	8.00×10^{15}	-1.2
160	1.48×10^{16}	1.50×10^{15}	0.7
200	2.21×10^{16}	2.27×10^{16}	1.3
250	2.66×10^{16}	2.78×10^{16}	2.2
350	2.92×10^{16}	3.07×10^{16}	2.5

beneficial, therefore, to expend considerable effort improving the experiment and the theory so as to decrease the amount of systematic error. In Section V, efforts to improve the state of the theory will be discussed.

The difference between the fits in Figures 11 and 12 can be seen in the region near a carrier concentration of $10^{14} cm^{-3}$. The two level fit in Figure 11 is systematically lower than the data in that region, whereas the three level

fit in Figure 12 follows the data more closely. Clearly, the fit shown in Figure 12 is a better one physically (following the shape of the data more closely) as well as statistically (giving a significantly lower value of the reduced chisquare). Note that this conclusion depends strongly on the accuracy and the density of data points in that region.

3. PARAMETER VALUES

The ultimate measure of the quality of any fit is the value of the parameters produced by the fit. No matter how low the chisquare or how well the curve follows the twists and turns of the data, if the values of the fitting parameters are physically unreasonable, the fit must be rejected. One cause of poor results is systematic error. When systematic error is present in the data or the theory, the fitting program distorts the values of the parameters so as to achieve the closest possible alignment between the curve representing the charge balance equation and the erroneous data. For this reason, the curve in Figure 8 is not a factor of 10 below the data (which is too high by a factor of 10), and the penalty for achieving even the poor fit shown is the multiplication of the acceptor concentration by a factor greater than 20 and a reduction of the ionization energy by 7% (See Table 1).

A poor value for a parameter may also be produced in a

41

"mushy" fit where the chisquare value is insensitive to the value of the parameter. This may be seen in data such as that in Figure 7 where a third impurity was required for a good fit, but the ionization energy of the impurity was found to have a poor value because the fit to the data is not very sensitive to that ionization energy. The solution here is to perform the fit with the ionization energy fixed at a value determined in another way, e.g. from optical measurements. Fixing the value of a poorly determined parameter is a convenient way of improving "mushy" fits even when the fitted values are not so far off the mark. Alternatively, there may just not be enough noise free data to determine a certain parameter well. In this case the scatter must be reduced or more data is required, usually over a wider temperature range or more closely spaced in temperature, in order to delineate bends and inflections better, thereby increasing the sensitivity of the experiment.

Occasionally, a fit may produce an unlikely value of a parameter because the fitting routine has found an alternate minimum of chisquare in the parameter space. In the situation shown in Figures 9 and 10, the fitting function produced good fits with two quite different sets of parameters and the fitting program found two independent local minima of chisquare in the parameter space. In order to insure that the "true" minimum has been found, it is

42

necessary to search for alternative minima in the parameter space and then examine the quality of the fit that each produces. Often, one fit is preferred because it gives a lower chisquare, because it follows the data better, or because it produces more physically realistic values of the parameters. For the data in Figures 9 and 10, the preferred fit is the one with a slightly larger value of the reduced chisquare. Both fits follow the data equally well, but the preferred fit has parameter values that are more reasonable physically. Subsidiary minima may be eliminated by fixing the value of one or more parameters when their values are known or by performing a series of experiments (Reference 11).

Theoretical systematic error can cause intriguing correlations of parameter values. For example, if the Hall scattering factor is taken to be 1 when fitting Si:In data, the ionization energy of the indium acceptor will depend on the amount of compensation of the indium. That is, if the indium acceptor is uncompensated, the indium ionization energy will be determined from the intermediate temperature range where the slope of the carrier concentration curve is related to one-half of the ionization energy, and a value in the range 0.165 - 0.170 eV is found. If, on the other hand, the indium acceptor is partially compensated so that the graph shows the low temperature slope related to the full

43

ionization energy, the value found is in the range 0.153 -
0.159 eV (References 11 and 17). Use of the empirical Hall
scattering factor discussed in Section V.3.b as a correction
for indium doped samples improves this state of affairs by
reducing the indium ionization energy in samples with low
compensation. As a consequence, the concentration of indium
found by the fit is also reduced in agreement with the
concentration found by other techniques (References 17 and
18). Neglecting the Hall scattering factor can cause an
error in the indium concentration as large as 100% and an
error in the indium ionization energy of about 7%. The only
unambiguous indication of this error in the Hall data itself
is the unusually high value of the indium ionization energy
in some r = 1 fits. Without correcting for the effect of the
Hall scattering factor or having other data, it can be
difficult to tell just how far the computer-produced
parameter values are from the true parameter values.

If accurate values are known, the values of the
parameters produced in a computer fit can be a very sensitive
test of the quality of the data and/or the theory. For
example, in Figure 8 above, a factor of 10 error in the data
translated into a maximum of a 30% difference between the
curve and the data for the best fit, whereas the impurity
concentation was in error by a factor of 20 and the
ionization energy by 7%. Also in Figure 9, a 30% error in

the theory due to the Hall scattering factor was reduced to about a 5% difference between the fit and the data points for an error of about 7% in the energy and 100% in the indium concentration. The error in the indium concentration for this sample could be reduced if data in the exhaustion (high temperature) region could be added. When it is available, data in all three temperature regions overdetermine the parameters in the fit via Equations 19, 20 and 21. Systematic error is most often revealed more in terms of a larger value of the reduced chisquare and less in terms of parameter error.

SECTION IV

THE FITTING ROUTINES

The computer program used to fit the carrier concentration data makes use of the computer subprograms (CURFIT and MATINV) in the book by Bevington (Reference 15). It finds the minimum value of chisquare by searching for the point in the parameter space where the gradient of chisquare is zero.

The program may be understood most easily if the discussion is confined to the case in which only two parameters, a_1 and a_2, are varied in the fit. The N data points (x_i, y_i) are fit using the fitting function $f(a_1, a_2, x_i)$ and Equation 27 for chisquare,

$$\chi^2 = \sum_{i=1}^{N} [y_i - f(a_1, a_2, x_i)]^2 / \sigma_i^2 \qquad (31)$$

is minimized to find the best value of the parameters a_1 and a_2. The optimum value of these parameters occurs at the point in the parameter space where chisquare is a minimum. We denote this point by $(a_1(\min), a_2(\min))$. At the minimum, the derivative of chisquare with respect to the parameters is zero,

$$\left.\frac{\partial \chi^2}{\partial a_1}\right|_{a_1(min)} = 0 \qquad (32)$$

$$\left.\frac{\partial \chi^2}{\partial a_2}\right|_{a_2(min)} = 0,$$

whereas away from the minimum the gradient will not be zero, but will point uphill toward larger values of chisquare.

Sufficiently close to the minimum, the value of the derivative of chisquare may be approximated by the first two terms in a Taylor's series exapansion,

$$0 = \left.\frac{\partial \chi^2}{\partial a_k}\right|_{a_k(min)} \simeq \frac{\partial \chi^2}{\partial a_k} + \sum_{j=1}^{2} [a_j(min) - a_j]\frac{\partial^2 \chi^2}{\partial a_k \partial a_j}, \qquad (33)$$

where the first and second derivatives of chisquare have the form

$$\frac{\partial \chi^2}{\partial a_k} = -2\beta_k = -2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} [y_i - f(a_1, a_2, x_i)]\frac{\partial f}{\partial a_k} \qquad (34)$$

$$\frac{\partial^2 \chi^2}{\partial a_k \partial a_j} = 2\alpha_{kj} = 2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left\{ \frac{\partial f}{\partial a_k} \frac{\partial f}{\partial a_j} - [y_i - f(a_1, a_2, x_i)]\frac{\partial^2 f}{\partial a_k \partial a_j} \right\} \qquad (35)$$

Note that the vector β points downhill toward smaller values of chisquare in the parameter space and α is related to the curvature matrix. Substituting these definitions into Equation 33 brings it to the form

$$0 \simeq -2\beta_i + \sum_{j=1}^{2} [a_j(min) - a_j] 2\alpha_{ij} \qquad (36)$$

which may be solved for the optimum value of the parameters

$$a_j(min) \simeq a_j + \sum_{i=1}^{2} (\alpha^{-1})_{ij}\beta_i . \qquad (37)$$

Because this equation is approximate, it does not in fact produce the optimum value of the parameters given any starting value, but produces a value closer to the optimum than the starting value. The starting point (a_1, a_2) in the parameter space is guessed, and the poorer that guess the worse the result produced by Equation 37. And so, Equation 37 should be viewed as a prescription for stepping toward the minimum rather than directly to it.

$$a_j(new) = a_j(old) + \sum_{i=1}^{2} (\alpha^{-1})_{ij}\beta_i \qquad (38)$$

In practice, the equation is applied iteratively until the minimum value of chisquare is reached.

In order to speed up the program, a further approximation is made. The second derivative of chisquare with respect to the parameters, Equation 35, involves the second derivative of the fitting function with respect to the parameters. Since this term is multiplied by the difference between the data value and the value of the fitting function,

it may be neglected if the functional value is close to the data value. Then, the matrix takes the form:

$$\alpha_{jk} \simeq \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \frac{\partial f}{\partial a_j} \frac{\partial f}{\partial a_k} \tag{39}$$

and its diagonal elements are positive. This second approximation makes the result of Equation 38 even worse when the initial guess is far from the minimum. This makes no difference so long as Equation 38 steps closer to the minimum each time it is applied. Since the vector β points toward lower chisquare, Equation 38 will step downhill if the step size is small enough and if it is in the direction of β.

To insure that Equation 38 always steps downhill, the inverse of a modified curvature matrix, α'_{jk} defined as

$$\alpha'_{jk} = \begin{cases} (1 + \lambda)\alpha_{jk} & \text{for } j = k, \lambda > 0 \\ \alpha_{jk} & \text{for } j \neq k \end{cases} \tag{40}$$

is used in it. Multiplying the diagonal elements of α by a value greater than one has the effect of making the inverse matrix smaller. This decreases the step size. Since only the diagonal elements are affected by this factor, a large value of λ will cause the diagonal elements to dominate when taking the inverse. Then, the inverse of α' is just the

matrix of the reciprocal of the diagonal elements of α' and since its diagonal elements are positive, the diagonal elements of the inverse will be positive. This has the effect of making the step in the general direction of β.

Larger values of λ are used to decrease the step size and point the step in the direction of the maximum decrease in chisquare (usually used far from the minimum), whereas small values of λ effectively remove this modification from the equations, allowing accurate matrix calculation of the next step (usually used near a minimum). The value of λ is changed dynamically in the program as the fit proceeds. It is multiplied by 10 if the previous attempted step was unsuccesful in decreasing chisquare and the step is reattempted. If the previous step successfully decreased chisquare, the value of λ is divided by 10 for the next step.

The program begins by asking the operator for an initial value of the parameters. The program computes the gradient of chisquare at this point and attempts to step away from this point to a new point in the direction of reduced chisquare. If the value of chisquare at the new point is the same or larger than the value at the original point, the attempted step was unsuccessful (i.e. uphill) and λ is made larger (to modify the size and direction of the step) and the step from this point is tried again. If the value of

50

chisquare at the new point is lower than the value at the original point, the attempt step was successful (i.e. downhill) and λ is made smaller for the calculation of the next step. The program continues stepping downhill until the value of chisquare stops changing.

A possible form for the appearance of the parameter space for a two parameter fitting function is shown in Figure 13. The oval curves are constant chisquare contours. Note that the computer steps three times to find the minimum. Note also that there are two minima in the diagram. The presence of multiple minima occurs fairly often, and some effort must be made to find all of the minima. If more than one minimum gives a good fit, other information must be brought to bear in selecting one of them as the "correct" fit.

The sharpness of the minimum and its uniqueness is a function of the quality and amount of the data. Reducing the random error in the data sharpens the minimum thereby permitting a more precise determination of the variable parameters. Reducing the systematic error in the data also improves the fit and allows a more accurate determination of the parameters. Taking more closely spaced data or taking data over the widest possible range of the independent variable (temperature in our case) can increase the precision
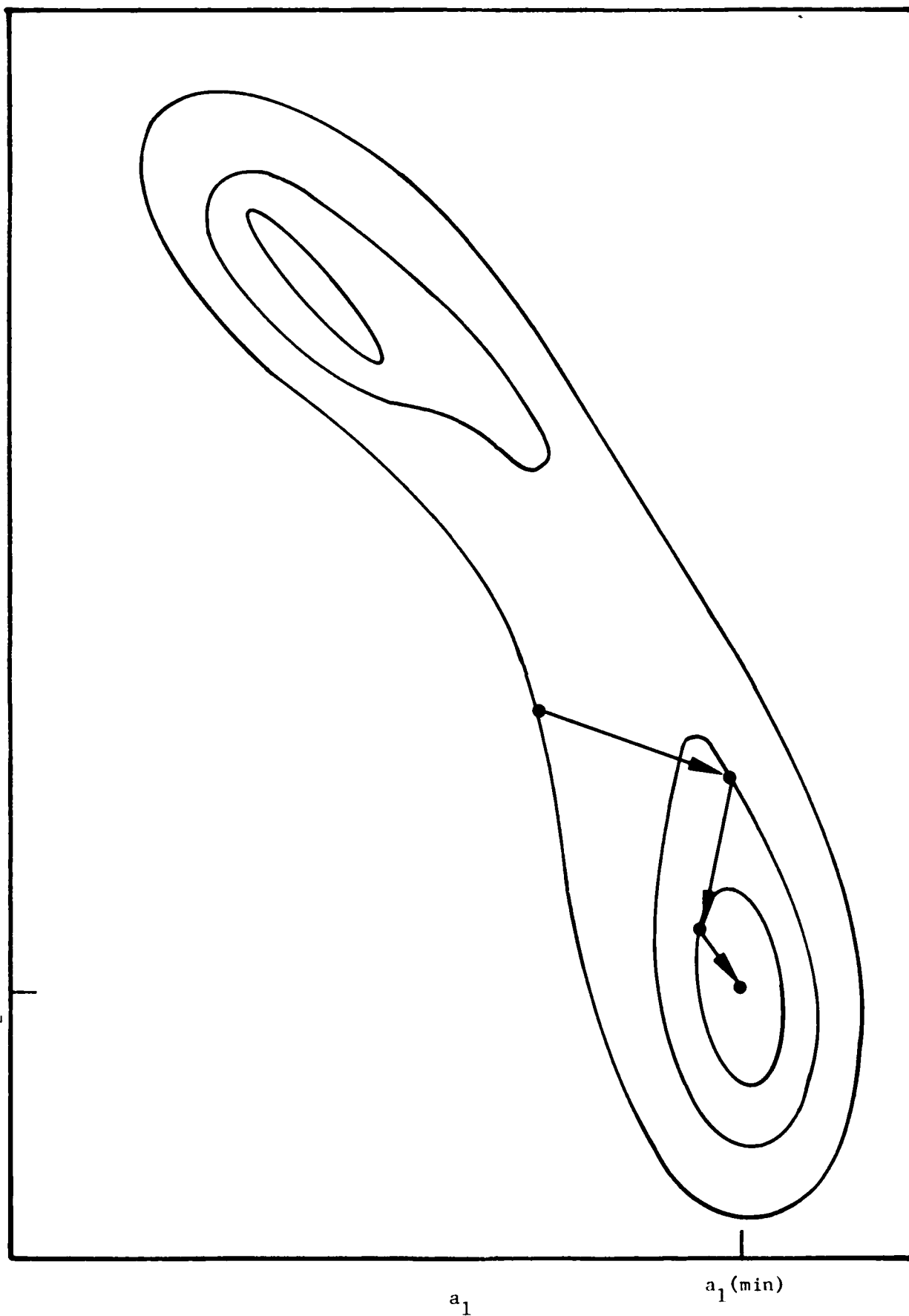
Figure 13. Contours of constants chisquare in a two dimensional parameter space. Two minima are shown in the diagram and the three arrows trace the steps taken to find the position of the minimum on the left.

a_2

a_2(min)

a_1

a_1(min)

of the results even in the presence of systematic error. All
of these strategies have the effect of increasing the amount
of information that may be extracted from the data via the
fitting program.

SECTION V

IMPROVEMENTS IN THE THEORY

For samples not too heavily doped, the true carrier concentration in the sample may be obtained accurately by solving the exact charge balance equation, but this is not usually done because of calculational difficulties. Instead the full equation is generally approximated by Equation 16. This is an excellent approximation for a doped sample and it only breaks down at high temperatures where mixed conduction becomes important. Also, the distribution of carriers in the bands is approximated using Boltzmann statistics, and ordinarily this too is a good approximation. In the first part of this section, the accuracy of these approximations will be estimated.

Besides the approximations, there are a multitude of constants in the charge balance equation which must be determined. We usually pick the concentrations (N_{Ai} and N_D) and ionization energy (E_{Ai}) of the impurities to be the parameters varied in the fitting process, and so the density of states in the band (N_V for the valence band), the degeneracy of the bound states of the acceptors (g_{ij}), and the binding energies of the excited states of the acceptors (E_{Aij}) must be put in by hand. In the second part of this section, these factors and the effect they have on the

54

calculated carrier concentration for p-type samples will be discussed.

Unfortunately, the Hall effect does not measure the true carrier concentration in the sample, but measures the carrier concentration divided by the Hall scattering factor, r. In order to find the true concentration of impurities in the sample the data must be corrected. The r-factor can be calculated theoretically which includes some uncertainties, it can be estimated by some measurements, and it can be measured explicitly in others. In the last part of this section, improved estimates of the r-factor will be assessed.

1.   APPROXIMATIONS

a.   Mixed Conduction

To obtain the modified charge balance equation (Equation 15) from the completely general form (Equation 14), the concentration of electrons in the conduction band must be neglected. This can be done if the concentration of conduction electrons is much less than the concentration of holes in the valence band. The maximum concentration of holes is given by

$$p = \sum_i N_{Ai} - N_D.$$ (41)

and the concentration of electrons is approximately

(References 2 and 7)

$$n = N_c N_v \exp(-E_g/kT)/p, \qquad (42)$$

where E_g is energy band gap and N_c is the density of states in the conduction band, so the condition for the neglect of the conduction electrons is

$$N_c N_v \exp(-E_g/kT)/p \ll p. \qquad (43)$$

At low temperatures this inequality is well satisfied because $E_g \gg E_{Ai}$ for any i, but as the temperature is increased conduction electrons become important first for lightly doped samples and only for heavily doped samples at the highest temperatures (Reference 19). In Figure 14, the Hall data for a high purity silicon sample is shown. Note the upward curve of the data at the left side of the graph. Here, the sample is making the transition to mixed conduction where both electrons and holes carry current. This upward curve of the data is indicative of the onset of mixed conduction and whereas Equations 5 and 15 may be used to obtain the carrier concentration at lower temperatures, Equations 1, 2 and 14 must be used in this high temperature region. In the absence of this type of evidence for mixed conduction, the modified charge balance equation describes the data very well.
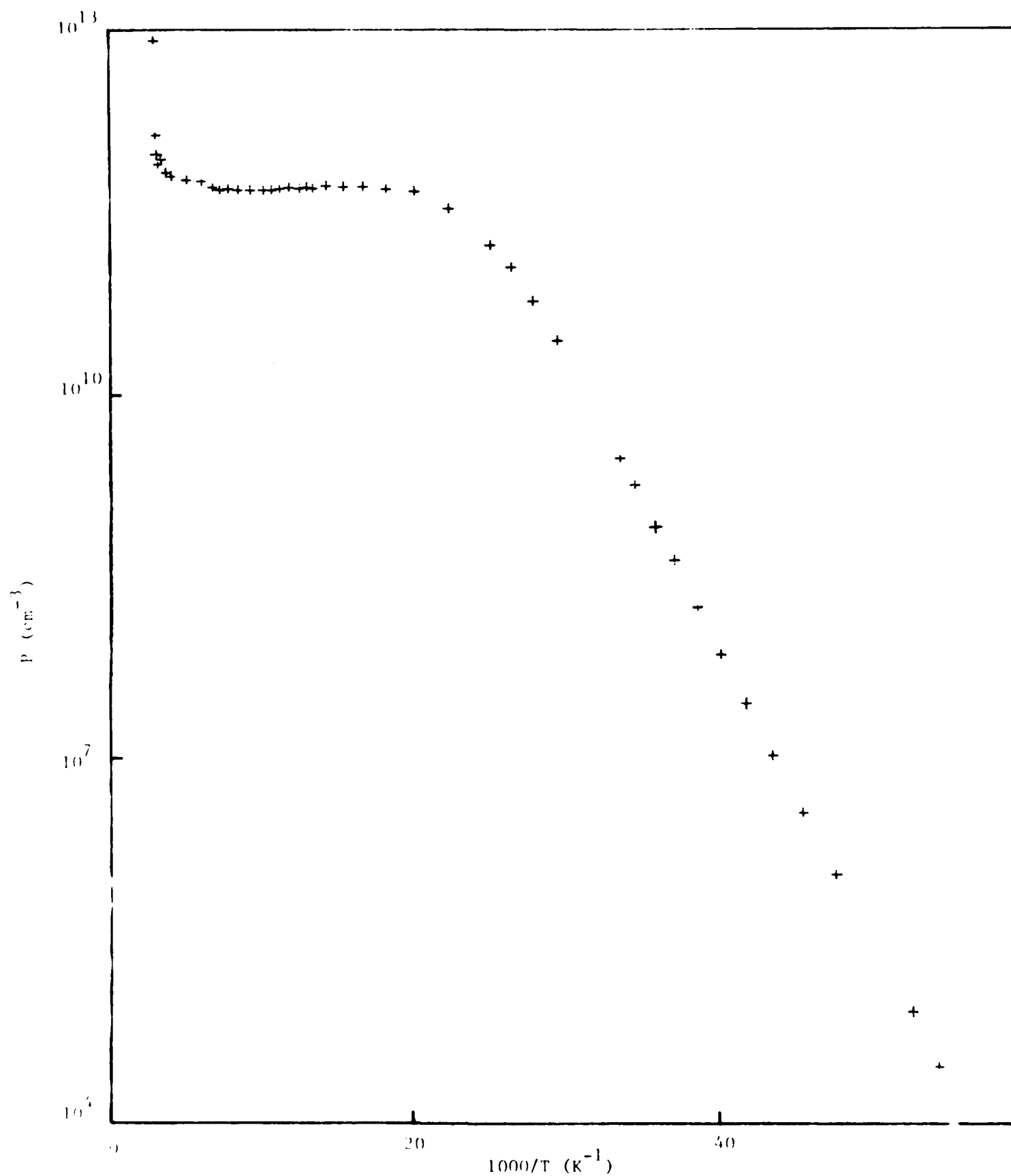
Figure 14. Carrier concentration versus 1000/T for a high purity p-type silicon sample. These data show the effect of the transition to intrinsic conduction at high temperature (left side of the graph).

b. Boltzmann Statistics

To obtain the form given in Equation 17 for the function Φ_i, the distribution of holes in the valence band was assumed to be described by Boltzmann statistics. Boltzmann statistics assumes a simple exponential distribution for the electron density and applies to electrons only if their density is low enough so that they are "nondegenerate" -- that is only if the Fermi-Dirac distribution function may be approximated by a simple exponential. This is a good approximation so long as the Fermi level, μ, is in the band gap more than 3.5kT from the edge of the valence band (Reference 7). In other words, the Fermi distribution in the integral is expanded in the form

$$[1+\exp((\mu-E)/kT)]^{-1} = \exp((E-\mu)/kT) \sum_{j=0}^{\infty} [-\exp((E-\mu)/kT)]^j \quad (44)$$

and the higher-order terms are neglected. This can be done so long as the absolute value of the sum of the higher-order terms is much much less than the first term:

$$\left| \sum_{j=1}^{\infty} [-\exp((E-\mu)/kT)]^j \right| \ll 1. \quad (45)$$

The maximum deviation from Boltzmann statistics will occur for holes with energies at the edge of the valence band, $E = 0$, and the most important term in the sum in Inequality 45 is the first one. For the condition stated by Blakemore,

58

$\mu = 3.5kT$, these terms represent at most a 3% correction at the edge of the band. Since the distribution function is integrated over energy, the error in the integral (which is the important number) is only about 1% (References 7 and 20).

This condition is hardest to satisfy for shallow dopants which can bring the Fermi level close to the band edge. At low temperatures, the Fermi level locks onto the ionization energy of the shallowest partially compensated acceptor. This is always many many times kT at these temperatures. At the highest temperatures, the Fermi level moves toward the middle of the band gap as the sample begins to experience mixed conduction and it again is many times kT from the edge of the band. It is only at intermediate temperatures that problems can develop when the Fermi level position relative to the band edge is one half of the ionization energy of the deepest acceptor. If the deepest acceptor is rather shallow, the Fermi level can approach the band edge to within a distance of kT. For a silicon sample heavily doped (density on the order of $10^{18}cm^{-3}$) with boron, a shallow acceptor, the condition $\mu \geq 3.5kT$ can break down if the exhaustion region is not entered by 73K. Actually, this requirement is much too stringent and the exhaustion region can be delayed to temperatures as high as room temperature before the error in the carrier concentration becomes greater than 1%.

2.   VALUES OF CONSTANTS

a. Density of States Effective Mass

The density of states in the valence band, N_V must be calculated and put into the charge balance equation. The only factor in it that must be handled with special care is the temperature dependent density of states effective mass. Recently it has been calculated for the valence band of silicon (Reference 21) and a formula has been published (Reference 22). Since the density of states is proportional to the temperature dependent density of states effective mass to the three-halves power, that factor is given in Table 3 and plotted in Figure 15 along with an earlier calculation by Barber (Reference 23). The greatest difference between the two curves occurs at the highest temperatures, but because the carrier concentration is not dependent on the density of states in the exhaustion region, the greatest effect on the carrier concentration occurs in the intermediate temperature region. The calculation of R. G. Humphreys (Reference 24) also shows a larger effective mass than that given by Barber.

Using a different effective mass, then, changes the shape of the carrier concentration curve. In comparison with the Barber mass, the new effective mass raises the calculated carrier concentration slightly at the upper end of the intermediate temperature region, this difference being greatest in samples which enter the exhaustion region at the highest temperatures. The effective mass now used for data

60
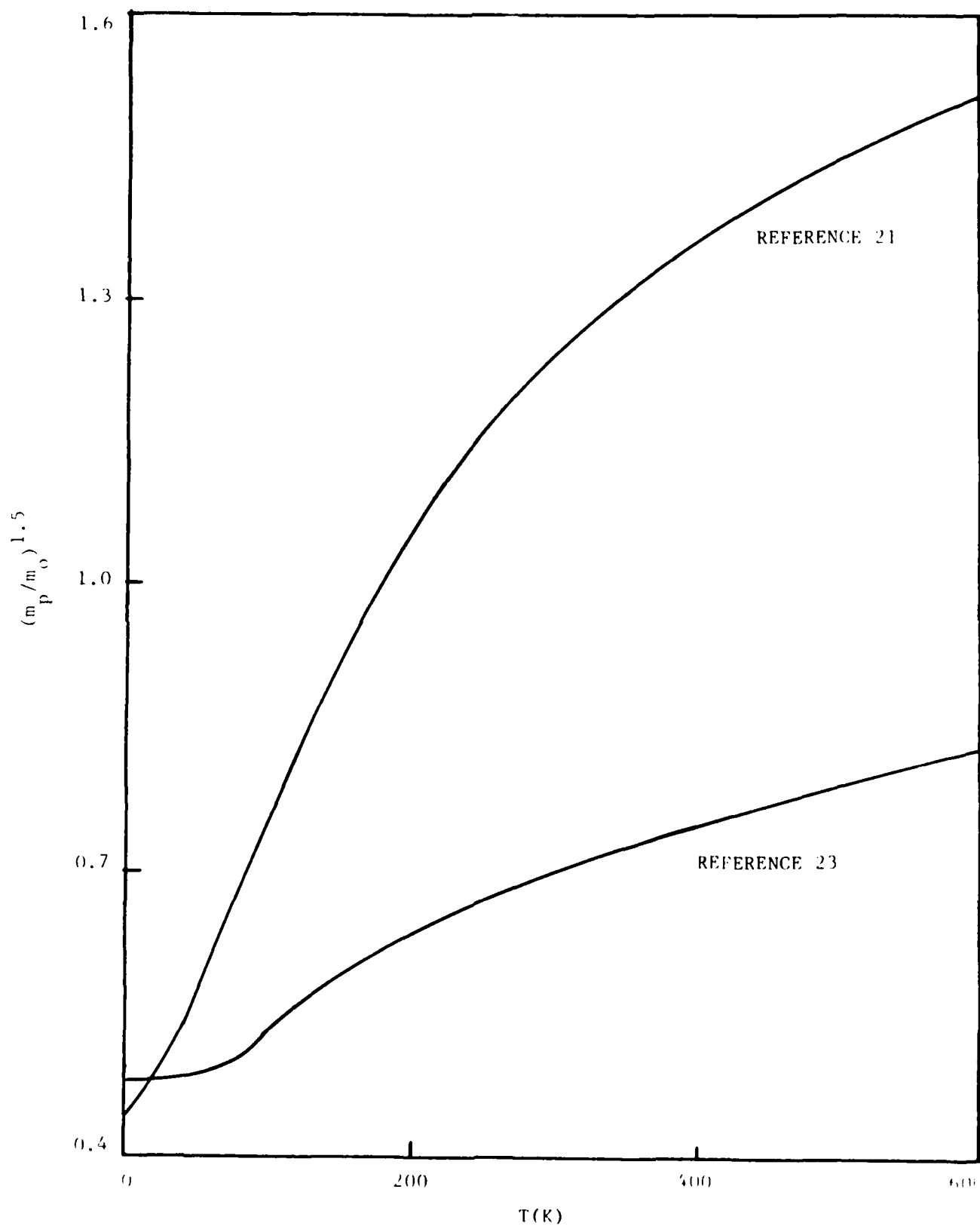
Figure 15. Density of states effective mass to the three-halves power versus temperature .

61

TABLE 3

DENSITY OF STATES EFFECTIVE MASS FACTOR

T (K)	$(m_p/m_o)^{1.5}$	T (K)	$(m_p/m_o)^{1.5}$
0	0.444	210	1.072
10	0.460	220	1.094
20	0.481	230	1.115
30	0.507	240	1.134
40	0.538	250	1.153
50	0.573	260	1.171
60	0.610	270	1.189
70	0.647	280	1.206
80	0.685	290	1.222
90	0.722	300	1.237
100	0.758	310	1.251
110	0.793	320	1.266
120	0.827	330	1.279
130	0.859	340	1.292
140	0.890	350	1.304
150	0.920	360	1.317
160	0.948	370	1.328
170	0.975	380	1.340
180	1.001	390	1.350
190	1.026	400	1.361
200	1.050	410	1.371

fitting is well-known and reliable, and no further refinement appears to be necessary at this time since it is not limiting the accuracy of the analysis.

b.   Excited States and Degeneracy

The excited states of the impurity contribute to the denominator of the formula for the function Φ_i, Equation 17. Their contribution may be assessed by rewriting Equation 17 in terms of an effective temperature dependent degeneracy factor, $G_i(T)$, in analogy with Equation 18:

$$\Phi_i = N_v \exp(-E_{Ai}/kT)/G_i(T). \tag{46}$$

Then, the effective degeneracy factor may be computed from

$$G_i(T) = \sum_j g_{ij} \exp((E_{Aij} - E_{Ai0})/kT), \tag{47}$$

where the sum is taken over all of the bound states of the impurity (including the ground state, $j = 0$) and $E_{Ai} = E_{Ai0}$ is the energy of the ground state and E_{Aij} is the energy of the jth excited state.

To apply Equation 47 to the case of the group-III acceptors in silicon, the g_{ij} and E_{Aij} must be known for all of the bound states for each acceptor. For the ground state, $j = 0$, the degeneracy is taken to be 4 and the binding energy is taken from the experiment of Fischer and Rome (Reference 25). For $j > 0$, the excited state energies and degeneracies are assumed to be the same for all acceptors, and the boron energy values (Reference 25) are used because the carrier concentration in boron doped silicon samples is most affected by the presence of the excited states. Degeneracies were assigned using the results of Lipari et. al. (Reference 26) and using relative intensity information (Reference 27).

The effective degeneracy factor is plotted in Figure 16 as a function of temperature for the four acceptors: boron,
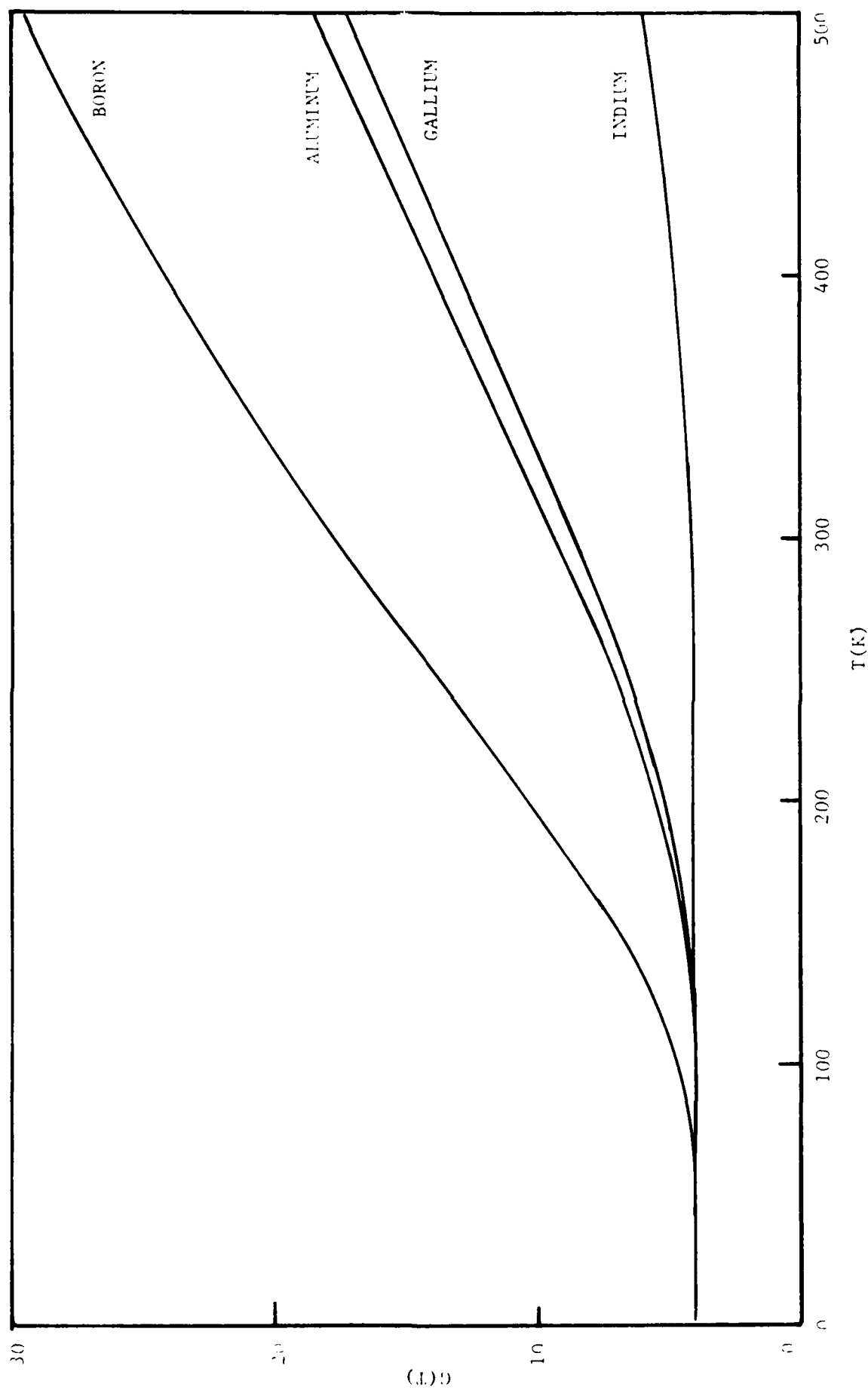
Figure 16. Temperature-dependent effective degeneracy factor for acceptors in silicon showing the effect of including the 22 known odd parity excited states of boron.
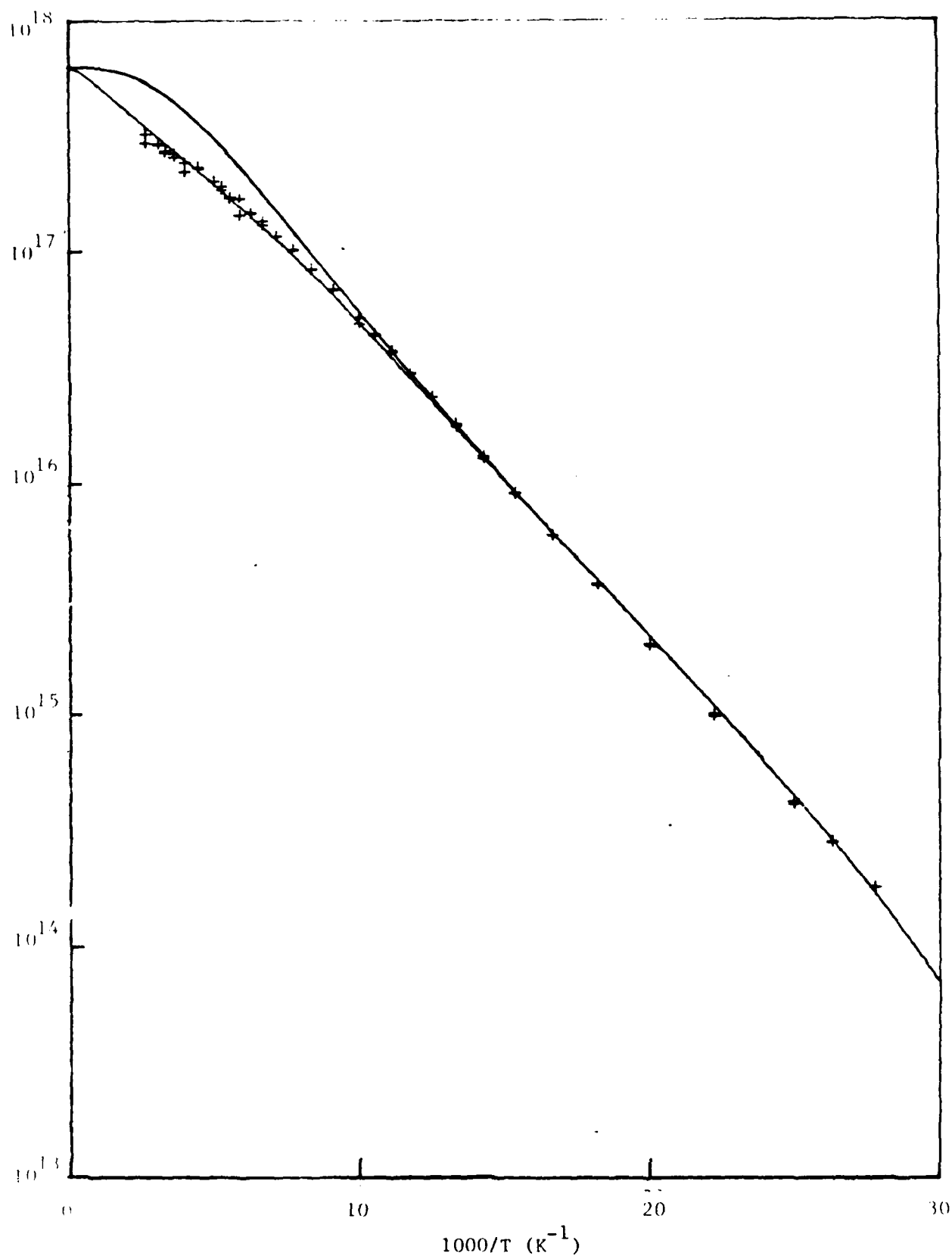
Figure 17. Carrier concentration data versus 1000/T for a silicon sample heavily doped with boron. The lower curve was obtained by fitting the data using a temperature dependent effective degeneracy factor that included 22 odd parity excited states and 5 even parity excited states. The upper curve was calculated with the same boron concentration ($6.25 \times 10^{17}/cm^3$) as for the lower curve but with a fixed degeneracy factor of 4.

aluminum, gallium, and indium. The ground state and 22 odd parity excited states reported by Fischer and Rome were included in the sum on the right hand side of Equation 47. Note that the effect of the excited states is greatest for the shallowest acceptor (boron) because they are closest to the ground state in boron.

In Figure 17, two curves are shown on the same graph as data taken on a boron doped sample. Both curves are calculated assuming the same impurity concentrations, but the lower curve includes the effect of 27 excited states (22 odd parity and 5 even parity corrected to the ionization energies of Fischer and Rome (References 26 and 28)). Note that the effect of the excited states is to reduce the calculated carrier concentration at the highest temperatures with the greatest percent effect occurring near the bend that separates the exhaustion region from the intermediate temperature region. Because this bend shifts to higher temperatures as the impurity concentration is increased, the effect of the excited states is most important for heavily doped samples.

The excited states now known and used give an effective degeneracy that is completely adequate for all but very heavily doped samples such as the one shown in Figure 17. Inclusion of excited states is essential for the reasonably

good fit in Figure 17 but the data in this figure need more excited states if the fit is to be improved further. Additional experimental measurements are therefore required to reveal the complete excited state structure, such as higher resolution optical absorption, photo-thermal ionization and Raman scattering to identify the even parity states.

3. ESTIMATES OF THE HALL SCATTERING FACTOR

The approximation $r = 1$ for the Hall scattering factor (or r-factor) is known to be a poor one because fits using it give incorrect values for the impurity ionization energy (Reference 17) and the impurity concentration (References 17 and 18). Since the value of r depends on the dominant scattering mechanisms in the sample, high purity samples, whose carrier mobility is dominated by acoustic and optical phonon scattering, will all have the same Hall scattering factor as a function of temperature whereas highly doped samples will have r-factors different from high purity samples and different from one another because their mobility is strongly influenced by ionized and neutral impurity scattering as well.

a. High-Purity Hall Scattering Factor

The earliest estimate of the high-purity r-factor was provided by Barry A. Green (Reference 9)) and was based on

his supposition that the temperature dependence of the carrier concentration of high-purity samples in the exhaustion region is caused by the temperature dependence of the r-factor. His estimate,

$$r = 0.83 + 0.04(T/1K)^{0.6}, \qquad (48)$$

was designed to have the proper temperature dependence in the range from 50 to 100 K, and therefore the first constant in the formula, 0.83, is only approximate. In Figure 18, this r-factor is plotted along with results of calculations by Szmulowicz and Madarasz (Reference 30) which are shown for comparison. These calculations include the scattering effects of only the acoustic phonons which dominate the mobility in high purity samples up to about 100K.

A recent experimental estimate of the high purity r-factor has been provided by Mitchel and Hemenger (Reference 31) who fit their data on the r-factor in the range 20-50K. They give the formula

$$r = 0.35(T/1K)^{0.082}, \qquad (49)$$

which is plotted in Figure 19 over an extended temperature range. Once again, the calculated r-factor of Szmulowicz and Madarasz is shown for comparison. These estimates show not only an increase of the Hall scattering factor in the region
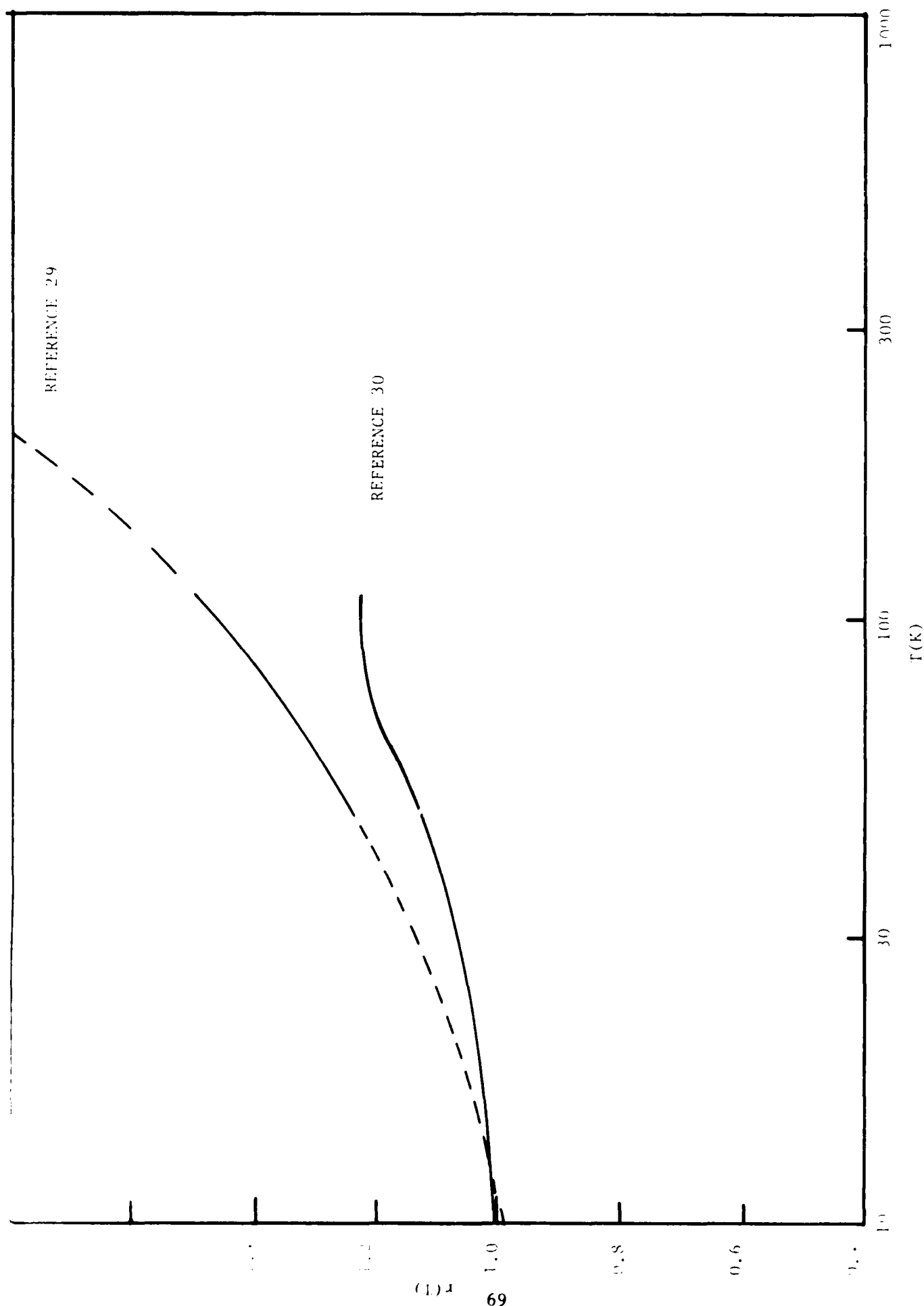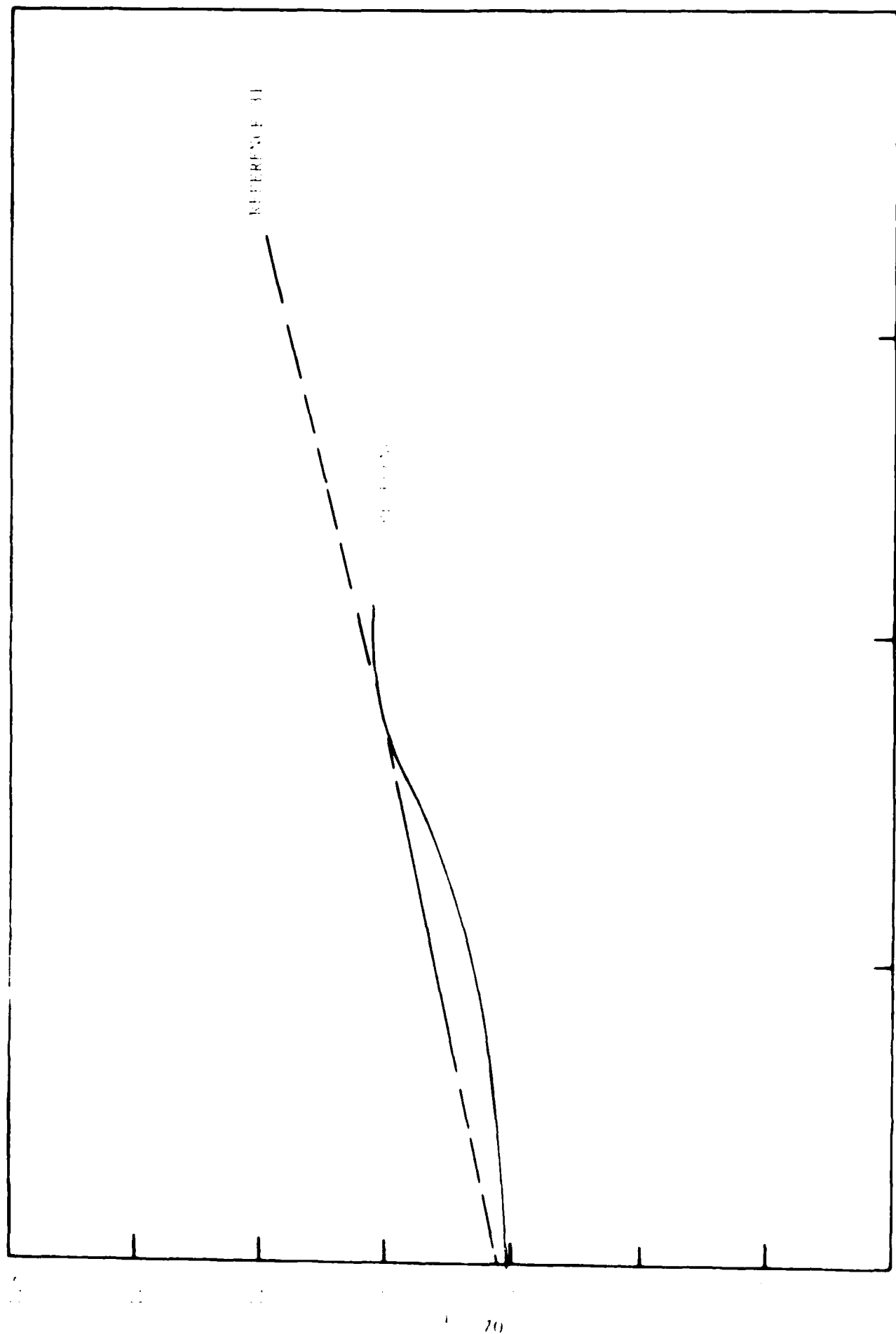
Figure 18. Estimated high purity r-factor of Barry A. Green (upper curve) and the calculated r-factor of Szmulowicz and Madarasz (lower curve).
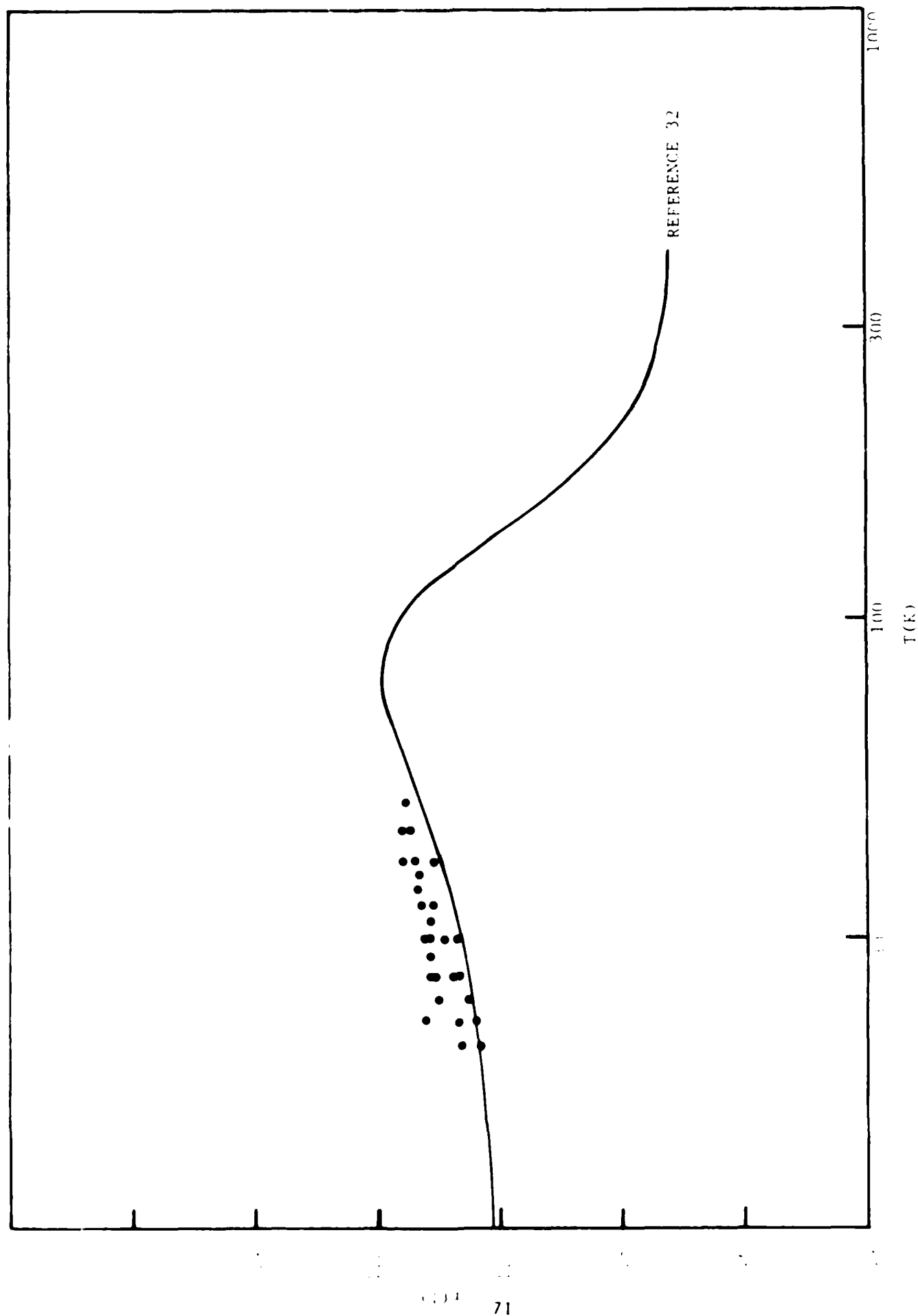
Figure 20. Data of Mitchel and Hemenger on the r-factor for p-type
silicon plotted along with the calculated r-factor of

from 20 to 100K, but taken together, also indicate a possible upward curvature in the data.

More recent calculations by Szmulowicz (Reference 32) includes optical phonon scattering in addition to acoustic phonon scattering. The r-factor calculated by Szmulowicz is shown in Figure 20 along with the data from Mitchel and Hemenger. This r-factor increases with T below 80K and also has the upward curvature from 20 to 80K, is about 4% below the data of Mitchel and Hemenger, and shows a peak below 100K. The peak in this r-factor at about 80K is inconsistent with the earlier estimate of Green which showed no peak within the range of his calculations up to 100K, but it is consistent with the calculation of Nakagawa and Zukotynski (Reference 33) which produces a peak near the same temperature. Szmulowicz's calculation used one adjustable parameter, the deformation potential, d_o, which was chosen so as to give a room temperature Hall mobility of 370 cm^2/Vsec. Reduction of this parameter moves the peak to higher temperatures and raises the room temperature mobility. A reduction of d_o even to zero, however, will not move the peak very far above 100K while it raises the room temperature mobility to unrealistically large values (Reference 32).

The carrier concentration data of Mitchel and Hemenger on sample 1202-H was taken over a temperature range from 18

72

to 100K at a fixed magnetic field strength of 1193 gauss, and a computer fit of this data can give an indication of the correctness of the r-factor. Unfortunately, at 1193 gauss, the sample is in the high field limit ($\omega_c \tau \gg 1$) at low temperatures and is in the low field limit ($\omega_c \tau \ll 1$) at high temperatures. Thus, the r-factor

$$r = R_H(0)/R_H(\infty) \tag{50}$$

cannot be applied to these data unless they are first corrected to zero magnetic field strength. Besides the temperature series, data was taken over a range of field strengths on samples 1202-H and 4802-3H at fixed temperatures from 20K to 60K giving an estimate of the correction to be applied to the 1193 gauss data.

After the estimated correction was applied to the 1202-H temperature series from 18 to 100K, two r-factors were used to fit the data and the results are shown in Table 4. The Szmulowicz r-factor was parametrized using the formula

$$r = 0.71739 + 0.28429f(x)\exp(-1.221025x), \tag{51}$$

where

$$f(x) = (1 + 0.43366x + 1.3219x^2)/(1 - 0.97429x + 0.46293x^2)$$

TABLE 4

RESULTS FROM FITS TO DATA FROM SAMPLE 1202-H

R-Factor	27 Data Points 18 to 100K		26 Data Points 18 to 90K	
	E_B (eV)	Reduced Chisquare	E_B (eV)	Reduced Chisquare
Szmulowicz	0.04492	0.71	0.04491	0.61
r = 1	0.04480	0.64	0.04471	0.61

and

$$x = T/80.526634K.$$

It gives very slightly poorer results than r = 1 mainly because of its peak near 80K. If the highest temperature data point at 100K is eliminated, it gives a reduced chisquare as good as that produced by r = 1. Fits using the Szmulowicz r-factor might be improved if the peak could be moved to higher temperatures. On the other hand, the data may contain errors due to the experimental limitation of magnetic field strength that only permitted a full field series to be taken up to 60K, which is the region dominated by acoustic phonon scattering, and this result was extrapolated to "correct" the data up to 100K.

Our understanding of scattering and the temperature dependence of the r-factor has improved dramatically over the

74

last few years. The present situation still leaves some questions that are hard to answer due to having data over only a limited range and possibly some issues remaining in the theory, such as the exact value for the optical deformation potential. We do have, however, a temperature dependent r-factor that agrees with the directly measured values over the range 20-50K and correctly removes the structure in the carrier concentration curve in the exhaustion region. This represents significant progress.

b.   Doped Sample Hall Scattering Factors

In doped samples, the r-factor situation is not nearly as simple as it is in high purity samples. Unfortunately, the calculation of Nakagawa and Zukotynski (Reference 33) has not helped much because the results are not directly comparable to experiment. Their graphs are plotted for the case of a fixed (temperature independent) number of ionized impurities in the sample. Actually, there is a great variation in the number of ionized impurities in the sample as the temperature changes, and that variation is strongly affected by the nature and strength of doping as well as the compensation. As a result, each sample has a unique r-factor.

Early estimates of the r-factor for doped samples concentrated on indium doped silicon where the need was

75

greatest. In these samples, the exhaustion region is not reached at the highest temperatures regularly measured (300-400K). The concentration of indium found in the Hall analysis fits for a sample was an inferred quantity which could not be checked directly using the carrier concentration in the exhaustion region. R. Baron et al. (Reference 34) extended their measurements to much higher temperatures to see the exhaustion region and found that the concentration of indium inferred from the lower temperature data did not agree at all well with the value obtained directly from the exhaustion region. Furthermore, they used C-V measurements as a check on the indium concentration revealed in the exhaustion region and also found disagreement. They concluded that the r-factor had a value of about 0.7 at room temperature and that it must rise as the temperature is lowered. The slope of the r-factor may be estimated from the incorrect value of the ionization energy for indium determined from Hall analysis.

Larry Green (Reference 35) used the results of Baron et al. as well as his own calculations of the r-factor for ionized impurity, neutral impurity and phonon scattering to present an estimate of the r-factor for indium doped samples. We improved and parameterized this estimate to produce the "empirical r-factor"
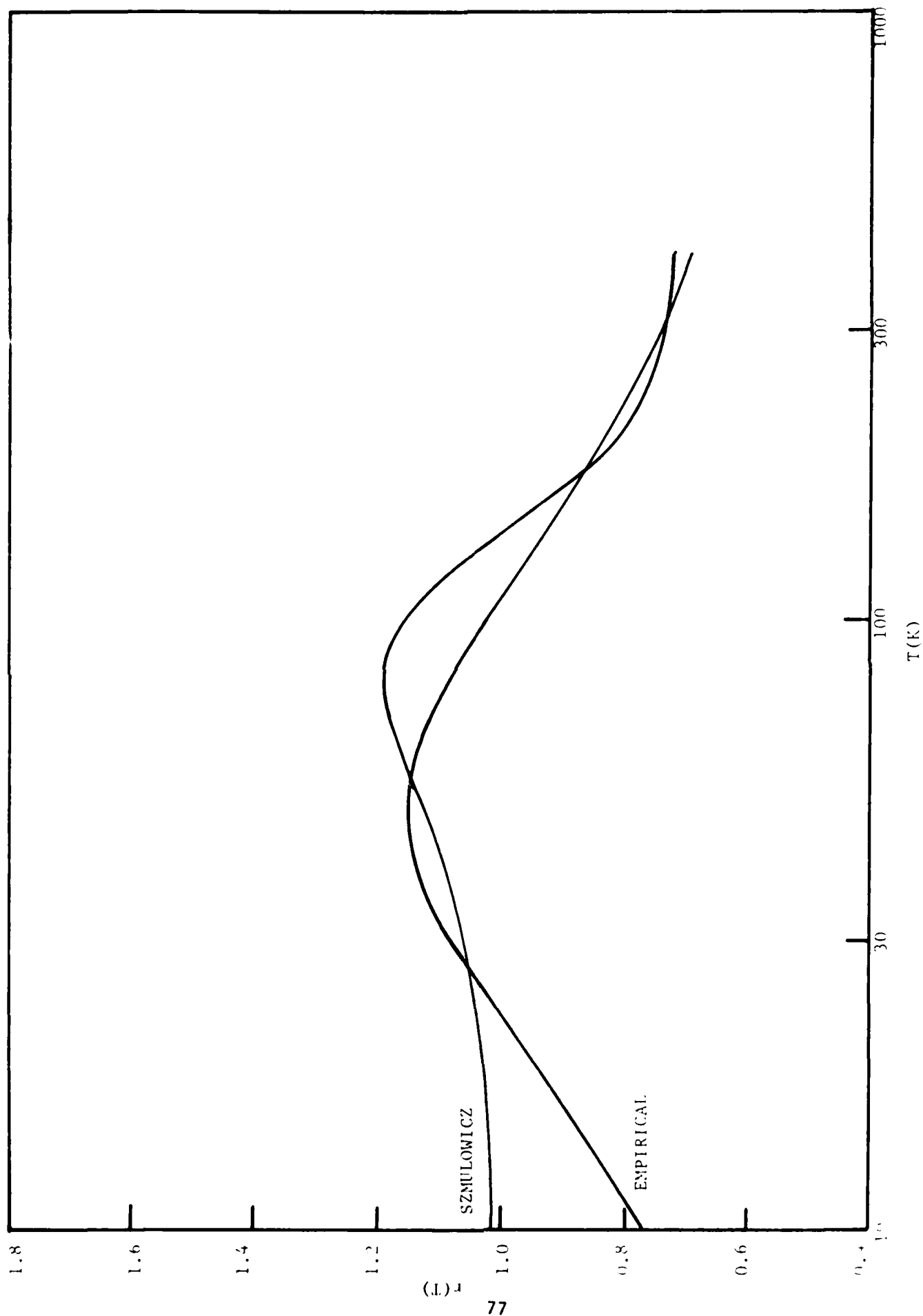
Figure 21. The "empirical r-factor" for indium doped silicon samples compared to the high purity r-factor calculated by Szmulowicz.

77

$$r = 0.54 + 1.22(T/50K)/(1 + (T/50K)^2). \qquad (52)$$

This empirical r-factor is plotted in Figure 21 along with the Szmulowicz high purity r-factor for comparison. Note that Equation 52 peaks at a lower temperature (50K) and has a more gentle high temperature slope than the high purity r-factor as a result of the rapid increase in the number of ionized impurity centers as the temperature is raised. This is consistent with the graphs of Nakagawa and Zukotynski (Reference 36).

A check on the empirical r-factor can be made by taking data on indium-doped silicon samples over a wider temperature range and seeing if the reduced indium concentration is consistent with the value of the carrier concentration in the exhaustion region. Indeed, high temperature data show a lower carrier concentration than that expected from a naive extrapolation of the low temperature r = 1 data (References 9 and 34), and the empirical r-factor produces fits to the low temperature data that are consistent with the high temperature data (Reference 9).

Two other r-factors have also been developed independently. The r-factor of Wucherer (Reference 37)

$$r = 0.61 + 0.489(1 - (T/300K)) \qquad (53)$$

was developed for gallium and indium doped samples. It lies below the empirical r-factor above 40K and has approximately the same slope between 120 and 200K. The r-factor of Holeman and Humphreys (Reference 38) is also a linear function of temperature. It is much lower than the empirical r-factor and its lower slope matches the slope of the empirical r-factor near room temperature.

Improved r-factors for doped samples must await a full calculation involving all scattering mechanisms because the high purity r-factor is only a rough guide to the value of the r-factor for doped samples. The conduction of both doped and high purity samples is dominated at high temperatures by optical phonon scattering and it is in that temperature range only where they could be expected to be similar. In Figure 20, note that the two curves are, in fact, very close to one another between 170 and 400K, but that they differ in value and shape below this range.

Recent calculations (Reference 39) indicate that the effects of ionized impurity scattering extend to much higher temperatures than expected and that, in samples with low compensation, neutral impurity scattering has considerably more effect at low temperatures than previously supposed.

REFERENCES

1. E. H. Hall, Amer. J. Math 2, 287 (1879).

2. E. H. Putley, The Hall Effect and Semi-conductor Physics, Dover Publications, New York, 1968.

3. Patrick M. Hemenger, Rev. Sci. Instrum. 44, 698 (1973).

4. L. J. van der Pauw, Phillips Res. Rep. 13, 1 (1958).

5. L. J. van der Pauw, Phillips Tech. Rev. 20, 220 (1958).

6. B. C. Dobbs, P. M. Hemenger, and S. R. Smith, J. Electronic Materials 6, 705 (1977).

7. J. S. Blakemore, Semiconductor Statistics, Pergamon, Oxford, 1962.

8. William C. Mitchel, private communication.

9. S. R. Smith, private communication of the low temperature data, and Gary E. Rattray, M.S. thesis, Air Force Institute of Technology, 1979, Report No. AFIT/GEP/PH/79D-9, for the high temperature data.

10. S. R. Smith, private communication.

11. R. D. Larrabee, J. Electrochem. Soc. 127, 1640 (1980).

12. M. C. Ohmer and J. E. Lang, Appl. Phys. Lett. 34, 750 (1979).

13. Jim Baukus, letter to P. M. Hemenger, Feb. 13, 1979.

14. P. M. Hemenger, J. E. Lang, V. Swaminathan and S. R. Smith, Bull. of APS 24, 276 (1979).

15. Philip R. Bevington, Data Reduction and Error Analysis for the Physical Sciences, McGraw-Hill, New York, 1969.

16. William C. Mitchel, private communication.

17. R. Baron, M. H. Young, J. K. Neeland, and O. J. Marsh, Appl. Phys. Lett. 30, 594 (1977).

18. D. K. Schroder, T. T. Braggins, and H. M. Hobgood, J. Appl. Phys. 49, 5256 (1978).

19. R. Baron, M. H. Young, J K. Neeland, and O. J. Marsh, Proc. of the 3rd International Symposium on Silicon

Materials Science and Technology (Silicon 77), Edited by
H. R. Hugg and E. Sirtl, The Electrochemical Society,
1977, p. 367.

20. W. B. Joyce and R. W. Dixon, Appl. Phys. Lett. 31, 354
(1977).

21. Frank L. Madarasz, Joseph E. Lang, and Patrick M.
Hemenger, J. Appl. Phys. 52, 4646 (1981).

22. Joseph E. Lang, Frank L. Madarasz, and Patrick M.
Hemenger, J. Appl. Phys. 54, 3612 (1983).

23. H. D. Barber, Solid-State Electron. 10, 1039 (1967).

24. R. G. Humphreys, J. Phys. C 14, 2935 (1981).

25. David W. Fischer and John J. Rome, Phys. Rev. B27, 4826
(1983).

26. N. O. Lipari, A. Baldereschi, M. L. W. Thewalt, Solid
State Comm. 33, 277 (1980).

27. David W. Fischer, private communication. (May 26, 1983).

28. Gail J. Brown and David W. Fischer, Solid State Comm.
53, 1145 (1985).

29. David Brown and Steven Smith, <u>Silicon Detector Materials
Characterization</u>, Air Force Materials Laboratory,
Wright-Patterson Air Force Base, Ohio 45433,
AFML-TR-78-152, November 1978.

30. Frank Szmulowicz and Frank L. Madarasz, Phys. Rev. B27,
6279 (1983).

31. A. C. Mitchel and P. M. Hemenger, J. Appl. Phys. 53,
6330 (1982).

32. Frank Szmulowicz, Phys. Rev. B23, 5943 (1983).

33. H. Nakagawa and S. Zukotynski, Can. J. Phys. 56, 364
(1978).

34. R. Baron, M. H. Young, J. P. Baukus, O. J. Marsh, and M.
J. Sheets, "Properties of Indium-Doped Silicon as an
Extrinsic Detector Material," <u>Meeting of the IRIS
Specialty Group on Infrared Detectors - Volume I</u>,
Infrared Information and Analysis Center, Environmental
REsearch Institute of Michigan, Ann Arbor, 1977.

35. Barry A. Green, "Hall Factor Effects on Determination of

81

Indium Concentrations in Si:In," March 27, 1979 (unpublished).

36. Frank Szmulowicz, private communication. (June 15, 1983).

37. Michael Wucherer, thesis, University of Erlangen-Nuremberg, 1981.

38. S. R. Holeman and R. G. Humphreys, "Analysis of Hall Data in Silicon," June 1981 (unpublished).

39. Frank Szmulowicz, Bull. Am. Phys. Soc. 30, 521 (1985).

END 2-87