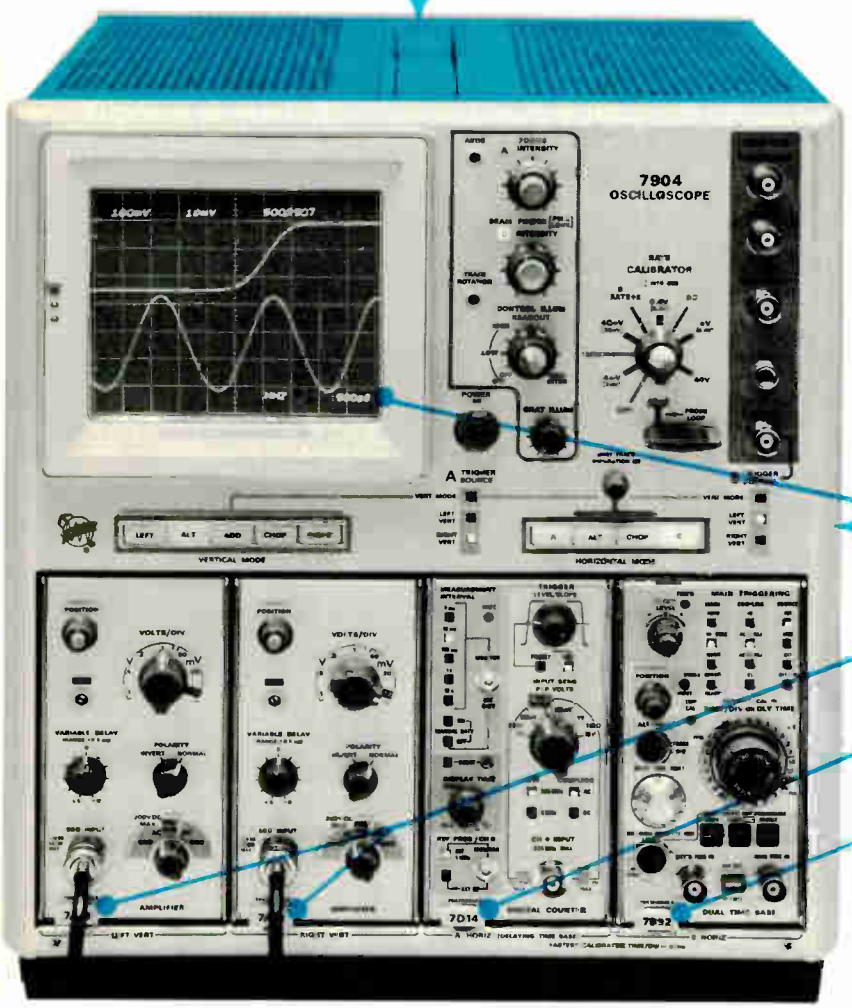
VOLUME 8 □ NUMBER 12 IEEE spectrum DECEMBER 1971

IEEE spectrum

features

TEKTRONIX® 7900 FAMILY

1 GHz Direct-Access Oscilloscope



TEKTRONIX 7904

featuring extended performance or general purpose use, from one mainframe and a family of plug-ins.

Extended Performance — 20 kHz to 1 GHz

Plug in the 7A21N and install a simple vertical amplifier bypass to directly access the CRT. The bandwidth — 1 GHz, and risetime — 350 ps. Less than 4 V/div driving signal required — single ended or differential. Sorry — no CRT READOUT — vertical amplifier bypassed.

General Purpose — DC to 500 MHz

Plug in the 7A19 — 500 MHz bandwidth at 10 mV/div, 7B92 — delaying sweep rates to 500 ps/div, 7D14 — direct counting to 525 MHz. As your applications and measurement requirements change, choose from 24 plug-ins for: • sampling • TDR • spectrum analysis • curve tracing • digital multimeter • etc., etc. TEKTRONIX 7904 . . . A product of technical excellence.

For further information or a demonstration call your nearby TEKTRONIX Field Engineer or write: Tektronix, Inc., P.O. Box 500, Beaverton, Oregon 97005.

Circle No. 2 on Reader Service Card

OR . . .

500 MHz real-time Oscilloscope System

8 x 10 cm display
CRT READOUT
9 cm/ns writing speed

7A19 Amplifier
500 MHz @ 10 mV/div

7D14 Digital Counter
525 MHz direct counting

7B92 Dual Time Base
500 ps/div

OR . . .

choose from 24 compatible 7000-Series plug-ins for virtually any measurement

7904 Oscilloscope	**$2900**
7904 Oscilloscope, without readout	**$2500**
7A19 Amplifier	**$500**
7A19 Amplifier, with variable delay	**$700**
7D14 Digital Counter	**$1400**
7B92 Dual Time Base	**$1400**
7A21N Direct Access	**$350**

U.S. Sales Prices FOB Beaverton, Oregon

Spectral lines

Our mission is venture. We have rejected the guild/AMA/trade union route to planned prosperity among a controlled population of electrical engineers. Let us assume that the proponents of that course do not make a sufficiently convincing case and we continue to reject it. Must we face the bleak future of an oversupplied labor commodity, exploited when employed and cast out in our mature years? We are in a unique position to avoid that pitiful fate for ourselves and the millions who manufacture, distribute, and service our creations because we can greatly enrich the quality of life for all of society.

The prospect that makes life good in our culture is the expectation of better things here on earth. It is becoming increasingly clear that the nature of those things must change from tons of metal and barrels of oil to subtler forms. With the exception of human affection, some of the subtlest, most kT-less forms of things for the race to use and enjoy are devised by electrical engineers. We are tooling brute labor and repetitive manipulation out of economic existence. The sweat of our brows will be reserved for tennis and gardening.

The economic work of humans is becoming the perception and processing of information. The physical work people once did is performed by machines instructed by a coded translation of their decisions. The rate at which this transformation occurs and its effectiveness depend directly on our imagination and enterprise today. We must open our minds and see things in new configurations to meet new functional specifications. Of course, gaps will appear in available technology but they provide the stimulation and selection criterions that the applied scientists need in their work.

There is nothing qualitatively new in what we must do. Technical innovation has always used what was available and developed new arts to meet a functional objective. What is needed now is a quantitative increase in the exploration of new potential applications. It seems like the wrong season for such a proposal. Government subsidies are drying up, companies are cutting every expense not clearly justified by a short-term business forecast. The soil is barren; it never rains and the business decision is to have the plowmen pack last year's crop. Well, we can do something effective about it. We have our minds and our mouths—in that order.

Our mission in this time of retrenchment is venture. We have the arts that will provide the work forms and recreation of the future. Let us take the initiative in finding the forms and modes in which they will be employed. One workable procedure is to start with an area with which we are familiar or can become familiar. Find the functions that are performed and see how they are performed. Question the functions deeply. What is the most basic statement of the service rendered? Is it likely to be needed for a long time? Propose alternate superior modes of rendering the service. Then, with this background, propose bold technological implementation of the superior modes. Another procedure, which is harder to initiate but offers greater rewards with success, is to search for services that are not yet in existence because they are inconceivable with older technology.

The proposal is that we all start this innovative process very deliberately. The probability of individual success is very poor, although the successful individual can be well rewarded, but this is not a lottery. If out of the 10^5 of us only 10 to 100 succeed, we all win. Our industry again will need more of us and the life of the world will be further enriched.

In today's business climate our proposals will have a mixed reception. Some businessmen will see them as a solution to their problems and others will be so dedicated to the conservation of liquidity that they will be rejected. If your efforts to promote your ideas fail internally you can propose protection and sale or protection and publication. You may choose to form your own business venture. In any case, the process of exposing your ideas to the serious consideration of others will reveal new ways of improving them.

The idea that we all seek new applications for our arts seems superficially obvious. Most of us feel we have been doing it all our lives and that is why we are engineers. The test for the validity of this proposal at this time is for each of us to ask himself whether he is working at it deliberately as though his future depended on it.

Good hunting!

David DeWitt, Editor

Computerization of English

Both error-correcting codes and natural languages exhibit a construction that alternates between the optional and the required. This principle is also the one underlying the construction of sentences by computer programs

Petr Beckmann *University of Colorado*

Natural languages, when regarded as codes for the transmission of information, are essentially structured like error-detecting or error-correcting codes. This principle has been used to write a program that, by random choice, is capable of constructing several billion different English sentences from a dictionary of less than a hundred unprocessed words. Applications to artificial intelligence and other fields are suggested.

Until recently, electrical engineering and linguistics had few common areas of interest. But the rapidly growing field of artificial intelligence not only is based on computer science; it also needs, among other things, insight into the structure of language. The close relationships between thinking, learning, and language are fully appreciated, though not fully understood, by psychologists. Linguistics, in its turn, has made great progress in the last two decades, especially in the development of generative and transformational grammars, among which that of Chomsky[1–3] is evidently the most widely accepted and highly developed. Yet Chomsky's grammar, even after a special computer language had been written for it,[4] attained only limited success as far as computerized sentence synthesis is concerned.

In the following, it will be attempted to show that much insight into the structure of natural languages can be obtained by using the principles of coding theory. In particular, it will be shown that the structure of natural languages is essentially that of an error-detecting (and, very often, of an error-correcting) code; when this is realized, it is a relatively simple matter to write a program that will produce grammatically correct, and even meaningful, English sentences from a dictionary containing words only in their basic forms. To support this contention, the author has written two sample programs. The first, when supplied with a dictionary of less than 100 words, is capable of constructing many billions of English sentences, all of them grammatically correct, but meaningless. The second program is grammatically simple, but it will produce English sentences (a possible total of about two billion from a dictionary of less than 100 words) that are not only grammatically correct, but also meaningful. Both programs make their decisions at random whenever English grammar leaves an option open.

The programs are actually simpler than the foregoing figures might perhaps suggest—which is borne out by the fact that they were written in Fortran (because no other compiler was available), a language quite unsuited for this type of work. A more appropriate source language,

such as Snobol, would probably reduce the present number of statements (600) by about two thirds.

Language as an error-detecting code

A natural language is a code for the transmission of information, and from this point of view is not radically different from a video signal, the Q-code used in aviation, or even the red–amber–green code of a traffic light. What is less obvious is the fact that the structures of error-detecting codes and natural languages show far-reaching analogies.

An error-detecting code, such as an (n, k) code, uses information and check digits. The information digits carry the message to be transmitted; the check digits check the information digits and themselves for possible errors.[5] In most computer codes, the check digits check the parity of the sums of certain digits; the Library of Congress Catalog Card Number code checks the remainder after weighting the digits and dividing the number by 11; etc. The actual mechanics are not important here. What is important is the principle: The user of the code is free to choose the information digits as he pleases; but, once he has done so, he must insert the corresponding check digits strictly by the rules of the code.

Natural languages appear to be structured on the same principle. They are not made up of digits, but morphemes. A morpheme is the smallest unit of grammatical significance; that is, a word that cannot be broken down further, such as *for*, *lid*, *hunt*, or a unit, such as *-able*, *-er*.

We are free to speak about a concept coded by, say, a noun—*usurer*, *urn*, *milk*, etc.—but the rules of English require us, quite redundantly, to make it clear immediately whether we mean a definite usurer or some usurer who has not been identified to the receiver (listener, reader) before. English grammar then requires a check morpheme, the article, to be inserted (or omitted, again by strict rules). Suppose the noun is to be indefinite. The rules of the code then require us to insert *a* before *usurer*—because its first phoneme is a consonant; *an* before *urn*—because its first phoneme is a vowel; and nothing before *milk*—because it is a collective noun.

The fact that the article is a redundant check morpheme is not only evident from telegrams and newspaper headlines (where it is omitted), but also from its nonexistence in other languages, such as Latin. The Slavic languages—except for Bulgarian—also do not have articles, yet the Slavs have no difficulty in communicating because these languages have an abundance of other check morphemes, in particular, inflections. (It is noteworthy that Bulgarian, the only Slavic language using articles, does not inflect.)

Gender is another criterion by which check morphemes are inserted. In English it is only weakly present (*he*, *she*, *it* and a few nouns, such as *waitress* and *actress*), and, except for ships and countries, it is determined entirely by sex. However, in French, German, Latin, and Russian, gender is usually independent of sex, or even in contradiction to it. The rules of these codes require check morphemes for the concord between noun and adjective (and also verb in Slavic languages), and obligatory check morphemes to form nouns that might be rendered in English as "workeress" or "driveress." The redundancy of gender is demonstrated by Hungarian, which does not know the concept: *he*, *she*, *it* are all expressed by the same word.

A third example of check morphemes is the ending prescribed in the conjugation of verbs, such as the third-person-singular ending *-s* in English (*can*, *may*, *must*, *ought* get on quite well without it). Just as a computer code asks for parity to insert its check digits whereas the Library of Congress code asks for divisibility by 11, so different natural languages have very different criterions for conjugating their verbs and choosing their tenses, i.e., criterions for performing the required checks. In English, the use of past or perfect is determined by the relation to the present: "I was here two hours ago"—"I have been here for two hours." In the Slavic languages, the conjugation depends on whether or not the action is completed; for example, *to break* and *to crack* are formed from the same root. In Hungarian, it depends on whether or not the verb is followed by a direct object.

The basic purpose of an error-detecting code is to protect the signal from distortion. In radio communications, distortion is most often the result of thermal noise; in Library of Congress numbers, it is most often due to human error; and in natural languages, it is most often caused by ambiguity. Consider Chomsky's well-known example, "Flying planes can be dangerous," where *flying* can be interpreted as a gerund (meaning "the flying of") or as an adjective (as in "the Flying Dutchman"). It is evident that the ambiguity is entirely due to the lack of the check morpheme *-s*, which the defective verb *can* does not take in the third person. As soon as it is replaced by a regular verb, such as *appear*—and the *to* of the infinitive, which must be omitted after *can*, is restored—the check morpheme *-s* resolves the ambiguity by its presence or absence: "Flying planes appear(s) to be dangerous."

It should be pointed out that, for historical reasons, English suffers from, or is blessed with, a remarkable lack of check morphemes (it is the noise resistance that suffers and the efficiency that is blessed), and therefore ambiguities of this type occur more often than in other languages. When Chomsky's sentence is translated into Czech, for example, it is protected from ambiguity by no less than four check morphemes of gender and conjugation, not counting the fact that Czech has no gerund and *fly* is intransitive.

The many other analogies of natural languages and digital error-correcting codes are discussed elsewhere.[6] It should be pointed out, however, that checks are made not only by morphemes, but sometimes also by words and even tenses. For example, any prescribed sequence of tenses is a check feature: Since it is predetermined by the main tense, in some cases uniquely, the subordinate tense carries, by definition, little or no information. Moreover, the subordinate tense is often semantically absurd—"till death do us part" is present tense, although obviously death is expected in the future and not during the wedding ceremony; "till death has parted us" is grammatically correct, but the tense even more absurd.

To summarize, both error-correcting codes and natural languages exhibit a construction that alternates between the optional and the required. This principle of choice–check–choice–check also underlies the construction of sentences by computer programs.

Coding characteristics of natural languages

Samuel Morse knew no information theory, yet it was obvious to him that if his code was to be efficient, it had to assign the shortest code word to the most frequent source symbol (E) and the longest code word to the rarest (J, Z). However, language will probably never evolve

into an optimum Shannon–Fano code, because the "best" points in signal space (and, of course, their environments required for protection against noise) have already been taken up, centuries ago, by words that are rarely needed now. More modern words sometimes try to elbow in as acronyms (*laser*) or abbreviations (*'fridge*), but as a rule they do not make it. Examples are

(English)			
tribe	*sword*	*unemployment*	*radiation*
(French)			
tribu	*épais*	*chommage*	*rayonnement*
(German)			
Stamm	*Schwert*	*Arbeitslosigkeit*	*Strahlung*
(Czech)			
kmen	*meč*	*nezaměstnanost*	*záření*
(Russian)			
ПЛЕМЯ	МЕЧ	БЕЗРАБОТИЦА	ИЗЛУЧЕНИЕ

However, this is apparently the only case in which natural languages do not follow the rules that a competent code designer would use. Let us give a few examples.

A well-designed code will place the maximum distance in signal space between the code words whose confusion either has a high probability or carries a large penalty.[5] Natural languages, too, code similar concepts by very different sounds:

(English)	*yes, no*	*left, right*	*two, three*
(French)	*oui, non*	*gauche, droite*	*deux, trois*
(German)	*ja, nein*	*links, rechts*	*zwei,* drei**
(Czech)	*ano, ne*	*levý, pravý*	*dva, tři*
(Hungarian)	*igen, nem*	*bal, jobb*	*kettő, három*
(Russian)	ДА, НЕТ	ЛЕВЫЙ, ПРАВЫЙ	ДВА, ТРИ

Conversely, where no confusion threatens, a well-designed code will economize by using the same code word for several meanings. For example, Fortran uses comma, point, slash, and brackets for several different instructions distinguished by context. Similarly, natural languages use homonyms (*calf*, *club*, *to set*)—but they also go much further. In particular, English has built a veritable system on this principle. Not only does it often use the same word as noun, adjective, and verb (*iron; Iron Curtain; to iron*), but it systematically uses the same form of the verb for the past and for the past participle, and the same form for present participle, gerund, nominalized verb, and adjective (*has been growing*, *growing wheat is profitable*, *the growing of wheat*, *growing children*). Obviously, such economy could not survive distortion by ambiguity without a special device, especially since English does not inflect (except for the Saxon genitive), and its conjugations have only very flimsy check morphemes. This special device is a rather rigid word order, in which the meaning of a word is defined by its position in the sentence. For example, "Cain killed Abel" does not mean the same as "Abel killed Cain," and the sequence "Abel Cain killed" is assigned no meaning. In contrast, in Russian or Latin the three words can be permutated arbitrarily without very significant changes in meaning because the check morphemes of inflection always show who is the killer and who is the victim. The word-order artifact is a very masterful one since there is no plausible noise that will attack word order, and the efficiency of English is enhanced without unduly lowering its noise resistance.

It follows from Shannon's channel capacity theorem that no code can be made 100 percent efficient without being mutilated by noise. The quality of a code therefore can be judged by the compromise that it strikes between efficiency and redundancy. However, artificial codes are designed for specific types of noise and for assumed noise levels. A simple error-correcting code is useless if the channel gives rise to burst errors, and a burst-error-correcting code is very inefficient in a channel causing independent errors. To the author's knowledge, no one yet has designed a self-adaptive code that will adjust its redundancy to the type and level of the noise. But natural languages can do just that. If no ambiguity threatens, a language will omit strings of information morphemes and leave the receiver to restore them from the check morphemes. Conversely, in slang, dialect, and colloquialisms, it will violate the rules of the code to omit redundant check morphemes.* For example, the Latin phrase *inter pares alias* means "under otherwise equal conditions," but it only says "among equal other." The check morpheme *-as* shows that the missing noun is feminine, accusative, and plural, from which the receiver is expected to restore *conditiones*. The Slavic languages, which have more check morphemes than Latin, use the same device even more drastically. A famous aria in a Czech opera starts with the sentence, "Every [man] regards his [sweetheart] as [the] only [girl in the world]," but the words in brackets are omitted. They are restored by the receiver from the remaining words, which are provided with check morphemes of gender, number, and inflection. To a more limited extent this is possible in French and German (*la pauvre*, *die Arme*), and even in English: "Got a match?" The check morpheme *got* makes the information morphemes *have you* superfluous. The opposite case of omitting redundant check morphemes usually violates the rules of the code, so that we find this case in slang, colloquial speech, and dialects. For example, the check morpheme *had* is shortened in "I'd better go," and omitted in "I better go." The check morpheme *-s* is omitted in "he don't," and the forms *am not*, *is not*, *are not*, *have not*, *has not* are replaced by the highly efficient homonym *ain't*. In French slang, the redundant check morpheme *ne* is omitted in the constructions *ne . . . pas*, *ne . . . rien*, *ne . . . que*, etc. In Czech slang, no difference is made between masculine and neuter adjectives, and the vocative case is often replaced by the nominative (it has already died out in Russian, except when calling God). All slangs shorten and contract (*going to* → *gonna*), thus raising the coding efficiency.

The high coding efficiency of (correct) English is attained by a remarkable lack of check morphemes and a relatively inflexible word order. It is also attained by a very densely populated signal space. As one would expect from coding theory, the high efficiency of English makes it prone to errors, and this is easily confirmed.[7] Since the Slavic languages use signal space very lavishly—i.e., with wasteful distances between the code words—a misprint of a single letter will rarely destroy, let alone

* This would seem to be an exception; however, telephone and radio have forced the Germans to adopt *zwo* for *zwei* in telephoning and broadcasting to avoid confusion with *drei*.

* Occasionally, omission of check morphemes is permitted; e.g., *that* may be omitted when it is either a conjunction or a relative pronoun in the accusative ("he said he agreed"; "the car I bought").

change, the meaning of a statement. This is not true in English: "Your first cost is your least cost." "Orders received before noon shipped some day." "Grime among youth is growing." A language abounding in check morphemes can easily afford the loss of some of them; English cannot. The omission of the article and other check morphemes in newspaper headlines leads to ambiguities so frequently that a recent collection,[8] limited only to those that appeared in print and were funny as well, lists some 1800 of them. The ambiguity of "Street-walker hurt in business section" is, of course, purely semantic, and will translate into other languages. However, the ambiguity in "Lawyer says he'll have baby in court," is caused by the lacking check morpheme *the* before *baby*. The verb–noun–adjective economy causes ambiguities of the type, "Dealers will hear car talk Friday noon." The gerund–adjective–nominalized verb–participle economy causes ambiguities such as "Man refuses to give up biting dog." (The check morpheme *his* before *biting* would resolve it.) The lack of the check morphemes *is* and *are* (which Russian does not need) leads to ambiguities of the type, "French army cooks women."

The story of natural languages as codes for the transmission of information can be carried much further,[6] but it has already taken us too far from the computer generation of natural languages. However, the reader may be interested in the historical reasons for the high efficiency of English. It is a phenomenon well known to linguists that during periods when a nation has little "culture" its language will develop rapidly, because it is not impeded by the standardizing influence of education and literature. In the case of English, the rapid development set in after the Norman conquest in 1066, when the educated spoke—and impeded—French, whereas the Anglo-Saxon serfs, during the next two centuries, raised the coding efficiency of English by shedding its redundant check morphemes. They made a superb job of it; in particular, all inflections (except the Saxon genitive) were deleted, and the verb was streamlined to such an extent that today the so-called "irregular" verbs in English are far more regular than the "regular" verbs in French. By the time English was revived as a written language, it had emerged as the language with the simplest grammar—and the most preposterous spelling—in the Indo-European group. We have a curious souvenir of those times, the words for the domestic animals and their meats. The former are Germanic, the latter Romance:

(German)	(English)	(English)	(French)
Kuh	*cow*	*beef*	*boeuf*
Schaf	*sheep*	*mutton*	*mouton*
Schwein	*pig (swine)*	*pork*	*porc*
Kalb	*calf*	*veal*	*veaux*

The reason is that it was the Anglo-Saxon serf who tended the animals, but the French—or French-speaking—lord who ate the meat.[9]

Computer-generated English

The two preceding sections have attempted to support the contention that the structure of natural languages is that of error-detecting codes in which the optional strings of information morphemes alternate with the required check morphemes. If that is so, then it must be possible to incorporate these rules (the grammar of a natural language) into a computer program, and to let the program construct grammatically correct sentences of a natural language in much the same way as the insertion of check digits in a digital code can be computerized. To demonstrate that this is indeed possible, a sample program generating English sentences has been written. It chooses the information items (unprocessed words) at random from a dictionary, and adds the obligatory check morphemes as dictated by its previous choice. It will also alter the syntax of the remaining sentence under construction, if this is required by the word it has chosen. The program does not, by any means, include all possible English structures, but it includes enough to make it plausible that any other structure can be added by using the same system. There seems to be no danger of running out of storage with presently available computers, since the dictionary is not part of the program and need not be stored in the core memory. The program supplies articles, auxiliary verbs, conjunctions, and prepositions (and could easily, but does not now, process adjectives into adverbs); all other parts of speech are stored in the dictionary. Nouns, adjectives, and verbs are stored with codes identifying their grammatical properties. Adjectives have a code number 0 or 1, depending on whether the indefinite article that is to precede them is *a* or *an*. Nouns have a code indicating whether the noun is countable, collective, abstract, or proper; and another to direct the program to the correct formation of the plural (*-s*, *-es*, *-ies*, *-ves*, etc., or irregular, in which case the program is transferred to a table look-up, where it will find *teeth*, *children*, etc.). The third code indicates whether the noun takes *a* or *an* for the indefinite article. Other grammatical characteristics of noun, such as gender, are not needed in this sample version. Verbs are coded to indicate the formation of the past tense and whether they are transitive or intransitive, or what preposition they require. Provisions are also made for verbs such as *fight*, which can be transitive, intransitive, or take one of several prepositions—*against*, *for*, *over*, etc. Irregular verbs will send the program to a table look-up. Other parts of speech in the dictionary are not coded at present.

It should be pointed out that although most linguists agree that a sentence need not be meaningful to be grammatical, there is little agreement as to what constitutes a grammatical sentence. The definition used here is the following: A sentence is grammatically correct if it violates no explicit rules of traditional grammar and can be converted into a meaningful sentence by substituting words with identical grammatical properties. To give a somewhat drastic example, the sentence "The bird was belittled by jealousy" is considered grammatically correct, since it does not grammatically differ from the sentence "The boy was plagued by loneliness."

The program goes through a number of forks and decides by the value of a random number which branches to take. The probabilities assigned to the individual branches of the fork can be arbitrarily biased from the data cards, and for each fork separately.

The first such random decision decides whether the subject of a sentence is to be a pronoun or a noun. If a noun is chosen, it decides whether it is to be definite or indefinite. However, if the selected noun was coded as proper, the fork is bypassed and the program is transferred to the definite branch of the fork. The program then decides whether the noun is to be singular or plural.

Again, if the noun is coded as collective, abstract, or proper (i.e., noncountable), the program is transferred to the singular branch without consulting the random routine. It then decides whether to take in a possessive adjective, and, if so, it chooses one—always at random. If the previous choices leave open the option of a numeral, the program is given the choice of adding one. By now the choice between cardinal and ordinal numeral is no longer optional; for example, if the noun is definite singular, the numeral must be ordinal. Next the program (possibly) selects an adjective and (possibly) a comma plus a second adjective. It now has all the information needed to decide whether the subject is to be preceded by *a*, *an*, *the*, or no article. This is the most complicated part of the program, because the rules for the English article are quite complicated. The problem is solved by a system of flags inserted in the various branches of the program. The branches not taken by the program leave these flags (variables) in their intial state zero, but the flags passed by the program will change their values to one. The program then computes the correct article (or its absence) from the value of the flags and the information given by the codes of the words that have been selected. For example, if the possessive-adjective flag is "up," no article is inserted; if the noun is countable, singular, indefinite, and an adjective flag is up, the program will decide on *a* or *an*, depending on the code of the (first) adjective; etc. The program is now ready to concatenate the chosen or computed words and to insert them as the subject into the array for storing the sentence that is to be constructed; for example,

AN UGLY, WHITE CAT

It now either can choose a second subject or proceed to the verb. In the former case, it is given the choice of *and*, *or*, or *with*, which it will concatenate, "strike" all flags, and return to the beginning. The result might be

AN UGLY, WHITE CAT AND YOUR THREE SMALL LOAVES

(*Loaf* is processed into *loaves* in the plural branch.) A verb is now selected at random, a tense is chosen, and it is decided—always at random—whether to negate the verb. If, for example, the choice is past conditional, negated, the program will go through the branches that yield HAVE, NOT, and through the branch that forms the past participle from the infinitive. The three are concatenated as soon as they are formed. If the verb is transitive, the program goes back to the subject routine, which it now uses as an object routine. If the verb requires a preposition, the preposition indicated by the verb code is concatenated first, so that the result might be

AN UGLY, WHITE CAT AND YOUR THREE SMALL LOAVES WOULD NOT HAVE RELIED ON THE THIRD UNDERTAKER.

The program is now given a choice of whether to concatenate a period, print the sentence, and start the next sentence, or to concatenate a conjunction and run its course once more (with provisions preventing a loop). In the latter case, the result might be

AN UGLY, WHITE CAT AND YOUR THREE SMALL LOAVES WOULD NOT HAVE RELIED ON THE THIRD UNDERTAKER, REGARDLESS OF WHETHER THEODORE TALKED ABOUT MY TEN CHILDREN. (1)

The program then starts a new sentence, which might be quite short, such as I PERSPIRED. (The code of intransitive verbs does not give the program the choice of running through the object routine.) It is noteworthy that the verb following "regardless of whether" is never negated ("regardless of whether Theodore did not talk" is not grammatical), whereas the conjunctions *although*, *but*, *because*, etc., can be followed either by an affirmative or a negated verb. This is done by two simple statements: a flag and a conditional transfer. Thus, the concatenated word controls the options of the syntax of the remaining sentence under construction, which, presumably, is not easy to do with a generative grammar of Chomsky's type. The same principle can be used to incorporate a prescribed sequence of tenses or the rules of indirect speech; Fig. 1. There seems to be no substantial reason why the foregoing principles could not be used for any other English constructions, since the program can choose, and even generate, its own "base structures" (general, basic phrase structures).

Although the many billions of sentences that can be produced in this way from a dictionary of less than 100 words are grammatically correct by the foregoing definition, they are, of course, pure gibberish. Before we describe how the sentences can be made meaningful, we note that even in this state the program can be used for some pastimes; Figs. 2 and 3.

To make the sentences meaningful is, within a small

FIGURE 1. Programming of the rules of indirect speech. If the introductory verb is in the present, any tense can follow it; if in the past, the subordinate tense must be past, pluperfect, present conditional, or past conditional. This is accomplished by inserting a flag called W(16) in the past-tense branch, which is "raised" (changes its value from 0 to 1) when the program passes it with an introductory verb. If the flag is down, the program will reach fork 12 and choose any tense; if it is up, this fork is bypassed and the program transferred to fork 13 by means of the Fortran statement "IF (W(16).EQ.1) GO TO 1234," where 1234 is the statement number of fork 13. The program is now offered only the choice of tenses that indirect speech permits.

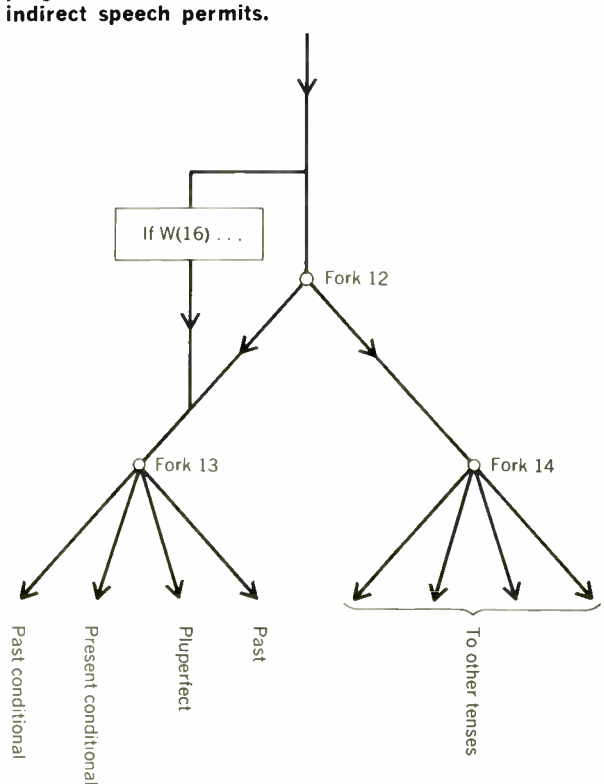


FIGURE 2. Meaningless alliterations. The dictionary has been loaded with words beginning with F or PH, pronouns and possessive adjectives have been excluded (by assigning zero probabilities to the corresponding branches of the forks). Conjunctions leading to a second run have been inactivated.

THE FRUSTRATED PHARAO AND THE FIFTH FLABBERGASTED PHOTOGRAPHER MIGHT HAVE FOOLED THE FILATELISTS.

A FOUL FROG SHOULD NOT FLATTER FABULOUS FORGERS.

FIVE FINS FUMED.

THE FORSAKEN, FEARFUL FRIGATE AND THE FORGETFUL FIRE FIGHTER FINISHED THE FIFTH FIREFLY.

FABULOUS FLUFF DID NOT FRATERNIZE WITH FATEFUL PHILIP.

FORTY-FIVE FANFARES AND A FIENDISH FOLDER MAY HAVE FANNED THE FIFTEENTH FINGER-LICKING PHARAO.

THE FRAULEIN MAY FRY A FIENDISH FIRE FIGHTER.

FIVE HUNDRED FRAULEINS OR A FORGETFUL FROG MAY FUME.

FRANCIS FORFEITED THE FINAL FRAUD.

FIVE FISHWIVES AND THE FIRE FIGHTERS FORNICATED WITH THE FRIVOLOUS FROGS.

THE FLABBERGASTED, FINGER-LICKING FRAULEINS DID NOT FUMBLE WITH A FLAMING PHARAO.

FRUITFUL, FASTIDIOUS FRIGATES DID NOT FOOL FOUR FROGS.

THE PHOTOGRAPHERS AND THE FOUR FINS WOULD NOT HAVE FULFILLED THE FLABBERGASTED FRIGATES.

THE FIFTIETH FAST FIRE FIGHTER MIGHT NOT HAVE FRIED THE FIVE HUNDRED FIENDISH FIRE FIGHTERS.

universe, simpler than to make them grammatical, because the rules governing meaning lend themselves more easily to computer processing than those governing grammar. We must take great care, however, to distinguish a false statement from a meaningless one. A moderately satisfactory rule is the following:

A meaningful statement must be either false or true; a statement is meaningless if it cannot be decided whether it is false or true. For example, "cats are green" is a meaningful statement since it can be shown to be false; however, the statement "cats are multiple-valued" is not false, but meaningless. If it were false, the statement "cats are single-valued" would be true.

To demonstrate the possibility of making the computer-generated sentences meaningful, the grammar was first simplified by excluding a second subject and conjunctions leading to a second run through the program, so that (1) would be reduced to

AN UGLY, WHITE CAT WOULD NOT HAVE RELIED ON THE THIRD UNDERTAKER

We can then satisfy the foregoing definition of meaning by preventing the following combinations:

1. An adjective qualifying a noun with respect to a property that the noun does not have ("the multiple-valued cat").
2. Contradictory adjectives ("the big, small cat").
3. A verb denoting an action that the subject is incapable of performing ("the cat diverges").
4. An object (or "quasi object") on which the action denoted by the verb cannot operate ("to differentiate a cat"; "to lean against the jealousy").

There are other meaningless combinations, such as the verb–adverb combination "to wilt hysterically," but they are not relevant to the program in its present state; they could be handled by the same principles.

The four rules are easily programmed. For example, to prevent contradictory adjectives, all mutually exclusive adjectives are coded by the same number. The program then multiplies the codes of the two adjectives to be tested for compatibility, and, if the product equals the square of one, it rejects the combination and selects a different adjective. A noun is coded for the properties with respect to which it can be modified; e.g., size, color, emotional state, intensity, numerical value, race, ... so that the

FIGURE 3. Spoof on legalese gobbledygook. The dictionary has been loaded with legal terms and nonexistent words. Possessive adjectives have been replaced by "said," ordinal numerals by "aforementioned," and the second subject is always joined by "and/or." In the verb branch, the program is given a choice of inserting a Latin phrase, and, if accepted, can choose from eight Latin phrases, which are concatenated, between commas, en bloc. This requires a small program modification. "Ad hoc" is stored as a regular adjective. The pronoun and cardinal numeral branches have been inactivated. No semantic coding has been used.

A LEGALISTIC, JUDICIAL DECISION AND/OR SUBSECTION B DID NOT, EX DEFINITIONE, OVERRULE A PRECEDENT.

SAID RETROSPECTIVE, OBSOLESCENT PRECEDENT AND/OR THE CITY ORDINANCE WULD EXONERATE A MISINTERPRETED DISTRICT COURT, REGARDLESS OF WHETHER SAID PRECEDENT AND/OR SAID TACIT, RETROACTIVE SUBSECTION REJECTED THE AFOREMENTIONED CITY ORDINANCE.

SAID PREJUDICIAL FEDERAL COURTS DID NOT EVADE PARAGRAPH 114/2, ALTHOUGH AN EXTRAPARENTAL PROCRASTINATION MAY NOT, SUB JUDICE, HAVE IGNORED THE INTERCOLLEGIATE SUBSECTIONS.

SAID BINDING LOWER DEFINITIONS AND/OR REACTIVATED PARAGRAPH 114/2 DID NOT REDEFINE THE AFOREMENTIONED INTRA-JUDICIAL, MISINTERPRETED TESTIMONIES, WHEREAS THE DISTRICT COURT DID NOT, AB INITIO, REJECT THE AD HOC, MISINTERPRETED VERDICTS.

AD HOC COUNTY COURTS AND/OR SAID INTRAJUDICIAL, BINDING CITY ORDINANCES REACTIVATED A PROCRASTINATION.

SAID JURISDICTIONAL APPEALS AND/OR THE BINDING PRECEDENTS DID NOT, EX DEFINITIONE, OVERRULE SAID DISTRICT COURTS, ALTHOUGH SECTION B WOULD EMPHASIZE THE CITY ORDINANCE.

THE AFOREMENTIONED ZONING LAW DID NOT IGNORE SECTION B, REGARDLESS OF WHETHER SAID RETROSPECTIVE, REACTIVATED SUBSECTIONS AND/OR SAID INJUNCTIONS SHOULD HAVE INTERPRETED THE INTRAJUDICIAL ZONING LAWS.

THE OBSOLESCENT DISTRICT COURT AND/OR THE INJUNCTIONS, POST FACTUM, REJECTED A RETROSPECTIVE, PREJUDICIAL PRECEDENT, WHEREAS SAID PARAGRAPHS AND/OR SAID SECTION B COULD NOT, IN LOCO PARENTIS, EXONERATE THE AFOREMENTIONED LOWER COUNTY COURT.

A VERDICT COULD OVERRULE SAID FEDERAL COURTS, REGARDLESS OF WHETHER SAID DEFINITIONS AND/OR A PARAGRAPH WOULD, AB INITIO, INVALIDATE SAID OBSOLESCENT, PREJUDICIAL ZONING LAWS.

THE AFOREMENTIONED COUNTY COURT COULD REDISESTABLISH SAID EXTRAPARENTAL SUBSECTION B, ALTHOUGH A LEGALISTIC TESTIMONY MAY NOT, SUB JUDICE, REACTIVATE RETROSPECTIVE PROCRASTINATIONS.

A RETROACTIVE DECISION COULD NOT, PRIMA FACIE, HAVE OVERRULED SAID TESTIMONY.

```
     THE TEN UNDERTAKERS DID NOT KILL MY CHIMPANZEES.
     AN UNDERTAKER MIGHT TALK ABOUT A TRANSCENDENTAL POLYNOMIAL.
     A CONSPICUOUS TOOTH MIGHT NOT HAVE KILLED A CHIMPANZEE.
     AVARICE SHOULD NOT ENRAGE THE NEXT SAD, BIG CHILD.
     A MALICIOUS CHILD VOMITED.
     CHILDREN MIGHT MATRICULATE.
     HIS FOURTH TRANSCENDENTAL DERIVATIVE MAY NOT HAVE VANISHED.
     A WAITRESS WOULD NOT HAVE FETCHED NINE LIBERALS.
     THE FIFTH LIBERAL TREMBLED.
     HIS MORONS SHOULD HAVE PLAYED WITH THE RED, WHITE
MEAT LOAF...****STOP****MEAT LOAF EITHER RED OR WHITE, BUT
NOT BOTH.
     THEY KILLED A MERRY, CONSPICUOUS DOG.
     HIS TEN IMAGINARY DERIVATIVES MIGHT HAVE DIVERGED.
     HER FIFTH DOG SHOULD NOT HAVE ENRAGED A CALF.
     YOU SHOULD NOT HAVE RELIED ON THE EXTREME LIBERALS.
     THE TWO TRANSCENDENTAL POLYNOMIALS DIVERGED.
     THE MEDIUM-SIZED CHIMPANZEE SHOULD NOT HAVE EATEN THE
VERY LIGHT, HEAVY ALUMINUM...****STOP****ALUMINUM EITHER
VERY LIGHT OR HEAVY, BUT NOT BOTH.
     EXTREME MELANCHOLY VANISHED.
     I VOMITED.
     THEIR DISCONTINUOUS POLYNOMIALS MAY NOT HAVE ENRAGED
THE PARANOIC CHILDREN.
     HER PARANOIC JEALOUSY ENRAGED THE TENTH SMALL WAITRESS.
     THEY WOULD HAVE EATEN THE WHITE MILK.
     WHITE, BLACK MILK...****STOP****MILK EITHER WHITE OR
BLACK, BUT NOT BOTH.
     THE WHITE DOGS SHOULD NOT SNEEZE AT A TOOTH.
     BIG CHILDREN DID NOT PLAY WITH SMALL THUMB TACKS.
     WE WOULD HAVE TALKED ABOUT HER EXTREME MELANCHOLY.
     THEY PERSPIRED.
     YOU DID NOT DIFFERENTIATE THE SINGLE-VALUED INTEGRAL.
     MY HEAVY, MERRY WAITRESSES DID NOT QUARREL WITH YOUR
PARANOIC OFFICERS.
     SHE DID NOT DIFFERENTIATE AN IRRATIONAL INTEGRAL.
```

```
     A MODERATE, IRRATIONAL LIBERAL DID NOT TALK ABOUT
A MEAT LOAF.
     OUR SINGLE-VALUED, DISCONTINUOUS INTEGRALS DID NOT
DIVERGE.
     I PLAYED WITH THE SIX CHILDREN.
     HIS SIX WHITE TEETH COULD NOT HAVE VANISHED.
     THE NEXT BLACK CALF COULD NOT HAVE SNEEZED AT WHITE MILK.
     I DID NOT DIFFERENTIATE MY IMAGINARY INTEGRALS.
     SHE HATED OUR FOUR POLYNOMIALS.
     THEY QUARRELED WITH MERRY, MALICIOUS OFFICERS.
     MY SMALL, BIG UNDERTAKERS...****STOP****UNDERTAKERS
EITHER SMALL OR BIG, BUT NOT BOTH.
     A HEAVY OFFICER SHOULD NOT RELY ON AN IRRATIONAL
UNDERTAKER.
     HE MAY HAVE DIED.
     OUR EXTREME LIBERALS WOULD NOT HAVE PERSPIRED.
     MORONS COULD NOT HAVE QUARRELED WITH THEIR CONSPICUOUS,
PARANOIC CHILDREN.
     MY UNDERTAKER MAY NOT HAVE VOMITED.
     FOUR MEDIUM-SIZED OFFICERS DID NOT DIFFERENTIATE A
POLYNOMIAL.
     SHE COULD EAT A MEAT LOAF.
     WE DID NOT FETCH THE VERY LIGHT THUMB TACK.
     YOUR IRRATIONAL AVARICE WILL HAVE VANISHED.
     A CALF MIGHT DIE.
     THE MALICIOUS, IRRATIONAL WAITRESSES ENRAGED THREE
BLACK CALVES.
     THE NEXT MALICIOUS CHILD SHOULD NOT HAVE EATEN THE
BLACK MEAT LOAF.
     THE IRRATIONAL POLYNOMIALS MAY HAVE DIVERGED.
```

code for *anarchist* would be 101101..., for *loneliness* 000100, for *waitress* 101001..., etc. The adjectives are similarly coded by the properties that they modify; for example, *extreme* 000100..., *black* 010001..., *irrational* 001010..., etc. The semantic noun and adjective codes are then unpacked, and corresponding digits multiplied; if no product differs from zero, the program goes back for another adjective. This will result in *irrational waitress* or *irrational function*, but not in *irrational aluminum;* in *extreme loneliness* or *extreme anarchist*, but not in *extreme waitress;* in *black waitress*, but not in *purple waitress* (since *black* is entered under both color and race, but *purple* under color only), etc. The verb is the easiest to code, since it need only be decided who or what can perform the action, and on whom or what it can operate. In the present program, the universe of nouns was subdivided into only five groups: persons, animals, tangible objects, emotional states, and mathematical functions. The verb *enrage*, for example, is then coded to take a noun from any group as a subject, but only a person or an animal as an object. In its present state, the program cannot handle a verb like *to wilt*, but could handle it if a subdivision of flowers were introduced (since *wilt* is intransitive, the absence of an object is already encoded in the grammatical code).

The dictionary supplied to the program contained 20 nouns, 20 adjectives, and 20 verbs, which corresponds to more than 1.28 billion sentences differing in semantic codes not counting variations in parts of speech, tense, negation, etc. However, since the program searches systematically, it will find the meaningful ones much faster than it can print—about two sentences a second. If two contradictory adjectives are selected (which is not too often), the program will not reject them outright, but will print the sentence up to the following noun, then interrupt the sentence and print the reason why it would be unacceptable; Fig. 4 shows part of the printout.

The difficulties of semantic coding grow rapidly with increasingly complex grammar and increasingly distant relationships, such as between different sentences. Although it is believed that the grammatical structure of the program can be enlarged without limitations, no such claim is made with respect to semantics. Consider, for example, the phrase "at either of the above addresses," which is meaningful only if the number of addresses "above" equals two. Although it is conceivable to write a subroutine that will search for, recognize, and count addresses in the previous output, it is obvious that an enormous number of such subroutines would be needed and that no computer could provide the storage capacity.

Applications

It is hoped that communication engineers and code designers may profit from an insight into the structure of language, as do aeronautical engineers from a study of the flight of birds.

FIGURE 4. Meaningful (but not necessarily true) statements. Since a polynomial must be finite and rational, the statement "The irrational polynomial may have diverged" is false; however, it is meaningful. If the program has selected two contradictory adjectives, it prints the sentence up to the next noun and then gives its reasons why the sentence would be unacceptable.

It is also hoped that insight into the structure of language from a new angle will prove helpful for linguists. The grammar used in the foregoing computer program is generative, but it does not contradict, refute, or improve Chomsky's transformational grammar. It simply appears better suited for certain applications, such as programming to generate both deep and surface structures, and for the interaction between them. It is, however, less general, and cannot be used for transformations or for detecting syntactic ambiguities. The principle of choice and check evidently also can be used for languages other than English, though the corresponding grammars will be different and more complicated.

The third hope is that the present approach will prove helpful in artificial intelligence, particularly for machine translation, language processing, question–answer systems, learning systems, and systems that might use a natural language as output. The points at which the program now makes its decision by consulting a random routine are accessible to the programmer (they can, in fact, be accessed as the arguments of the whole program running as a subroutine), so that the resulting subroutine could be made "to say in English" what is required by the input controlled by some other program. It would not, for example, require much effort to make the program, simply by inputting a string of numbers biasing the corresponding probability forks, produce any one of the billions of possible sentences. The same program could produce sentences as disparate as Shaw's statement:

DEMOCRACY SUBSTITUTES ELECTION BY THE INCOMPETENT MANY FOR APPOINTMENT BY THE CORRUPT FEW

and the ungrammatical statement:

I AIN'T GOT NO NUTTN'.

The latter statement is the more difficult one; the program would have to be "tricked" into producing the treble negative by entering *no* and *nuttn'* in the dictionary under "deceptive" codes to "smuggle" them past the corresponding flags.

In this and many other applications, the meaning is already implicitly contained in the input, so that the "gibberish" version of the present program could be used. On the other hand, there are applications, such as machine translation, where the semantic coding could open up new possibilities. For example, Russian homonyms, which are now machine-translated as *lock/castle*, *carry/wear*, or *face/person*, could be resolved; in fact, the *face/person* (ЛИЦО) homonym could probably be resolved in most contexts even by the sample program.

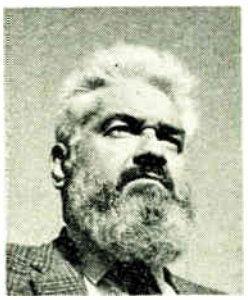
Finally, the choice-and-check approach also may prove useful for natural intelligence. No one knows how the brain constructs sentences, but the writer suspects that this method might be nearer the mark than transformational grammars. There is some evidence for this, though admittedly it is very flimsy. In learning a language, the student (a native child or a foreigner) learns to handle informational items much faster than the check system. In English, a child will produce nonredundant utterances such as "Annie want milk" or simplified check morphemes such as in "I eated." (In language with more check morphemes this is even more apparent.) People combine the check system of their native languages with the informational items of another language: A German will say "I am here since yesterday," a Frenchman will say "I love the nature," a Hungarian will speak of a woman as "he," and Slavs will leave out articles. In their endeavor to master the system both children and foreigners will look for a consistency that the system does not have. They will use (or ask for the meaning of) words like *couth* and *ruly* in the belief that they are opposites of *uncouth* and *unruly;* they will invent words that the languages "should" have, such as *allbody*, *eachbody*, *nephew-in-law*, *unhungry*, *to bejoy*, *onliness;* if shoes are *taken off*, they "should" also be *taken on;* if they are *put on*, they "should" also be *put off*. What this, and other arguments, might suggest is that the brain handles the informational structure of language differently from the check structure, and that it might construct sentences in some way feebly reminiscent of the choice-and-check sequences of the program described.

Whatever other applications this approach to language and the corresponding computer program might have remains to be seen, but I have no doubt about one: Language is the most fascinating toy ever devised.

REFERENCES

1. Chomsky, N., *Syntactic Structures*. The Hague–Paris: Mouton, 1957.
2. Chomsky, N., *Topics in the Theory of Generative Grammar*. The Hague–Paris: Mouton, 1966.
3. Chomsky, N., *Aspects of the Theory of Syntax*. Cambridge, Mass.: M.I.T. Press, 1965.
4. Yngve, V. H., *An Introduction to COMIT Programming*. Cambridge, Mass.: M.I.T. Press, 1966.
5. Beckmann, P., *Probability in Communication Engineering*. New York: Harcourt, 1967.
6. Beckmann, P., *The Structure of Language—A New Approach*. Boulder, Colo.: Golem Press (to be published, 1972).
7. Beckmann, P., Versuch einer semantischen Informationstheorie mit Anwendungen auf gesprochene Sprachen, *Wiss. Z. Hochsc. Elektrotech., Ilmenau*, vol. 4, no. 3, pp. 275–297, 1958.
8. Tempel, E., *Humor in the Headlines*. New York: Pocket Books, 1969.
9. Bradley H., *The Making of English*. New York: St. Martins Press, 1904.

Reprints of this article (X71-121) are available to readers. Please use the order form on page 8, which gives information and prices.



Petr Beckmann (SM) was born in Prague, Czechoslovakia. He obtained the M.Sc. (1949) and Ph.D. (1955) degrees in electrical engineering from Prague Technical University, and the Dr.Sc. degree (1961) from the Czechoslovak Academy of Sciences. He worked as a research scientist in the field of radiowave propagation at an institute of the Academy in Prague until 1963, when he was invited to the University of Colorado as a visiting professor, subsequently remaining there in his present post as professor of electrical engineering. Dr. Beckmann's main fields of interest are electromagnetics and applied probability, in which area he has published more than 50 papers and five books. Among his other interests are history (he is the author of "A History of π") and linguistics. In the latter field he has applied both information theory and computer programming, also drawing on his experiences as teacher of English, interpreter, and translator of scientific books.

An introduction to IC testing

Far from being an adjunct to the production of integrated circuits, stage-by-stage testing represents a vital, and highly complex, part of the manufacturing process

Frederick Van Veen *Teradyne, Inc.*

IC testing has evolved from the patterns established some years ago in the production of semiconductors. Since manual testing cannot meet the complex needs indigenous to IC manufacture, highly sophisticated instruments and test systems have developed that are automatically programmed by computer, tape, or printed-circuit card. This article focuses on many of the problems encountered and techniques employed, and also the requirements imposed on automatic IC testing systems.

More than 300 million monolithic integrated circuits will have been sold in the United States during 1971, and a conservative guess on industry-wide yield would lead one to think that well over a billion chips were tested in order to produce these 300 million marketable ICs. Each good circuit, in the course of its travel from wafer to end use, is probably tested an average of three times, and each test involves the qualification of many circuit functions and parameters.

Without carrying this exercise to extremes, it is obvious that the integrated circuit has brought with it a tidal wave of testing. Or, to turn the picture around, the introduction of automatic equipment capable of making tens of thousands of tests on a device in a few seconds has made the integrated circuit commercially viable.

Because IC testing is in the mainstream of the production process, it is very much in the mainstream of the chronic IC cost/price competition. For this reason the technology of IC testing is always colored by economic considerations. It is usually far more important to accommodate another multiplexed test station than it is to add another digit of resolution.

Evolution

IC testing is, of course, only a logical extension of transistor and diode testing, so we must look about ten years into the past for the origins of the patterns that have unfolded.

The first automatic semiconductor test equipment to be marketed commercially was manually programmed to supply the proper biases, limits, and classification criterions. Once programmed, the instruments were fully automatic in operation, the chief restriction on test rate being the speed with which an operator could insert transistors or diodes into the test socket and place them in bins according to test results. Automatic handling equipment was soon introduced to perform both of these functions, and multiplexers were developed to give the test equipment even greater leverage.

The manually programmed semiconductor tester survives in the form of instruments used for the incoming inspection and evaluation of transistors, diodes, and, to a very limited degree, ICs. It represents a practical approach to the problem of testing relatively small quantities of devices, each of which requires a relatively small number of tests. Most IC testing, however, involves complexity beyond the practical limitations of manual programming.

This situation has led to the development of the test equipment on which most attention is focused these days—instruments and test systems automatically programmed by computer, tape, or printed-circuit card. The computer, of course, offers not only test-plan storage but also the important ability to base the conditions of one test on the results of a previous test. The computer (and in particular the minicomputer) has become a major element in the IC testing picture, and it is typical of our bootstrap technology that the minicomputer is in a very real sense a product of its own making.

Figure 1 is a generalized clock diagram of a computer-controlled IC test system, as represented by a number of commercial equipments. The test program is loaded into the computer via paper or magnetic tape or by punched cards. Instructions from the computer are sent to an interface or control unit, which passes them along to the appropriate elements of the system.

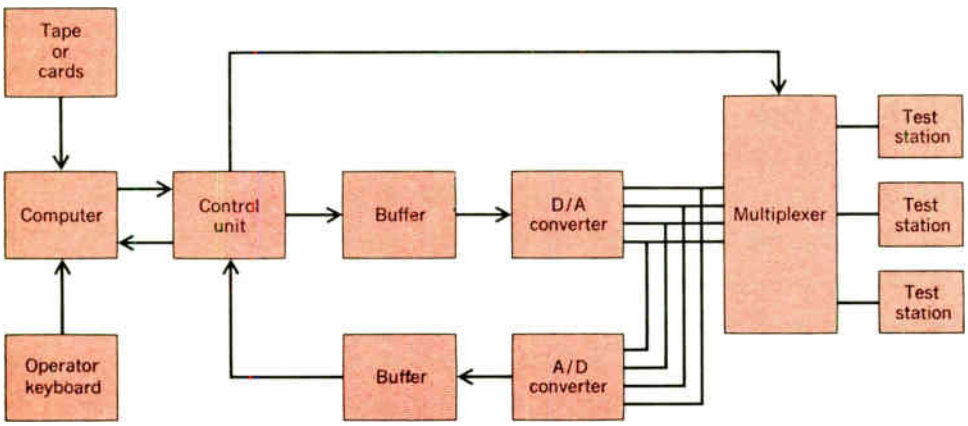


FIGURE 1. Basic configuration of a computer-controlled IC test system.

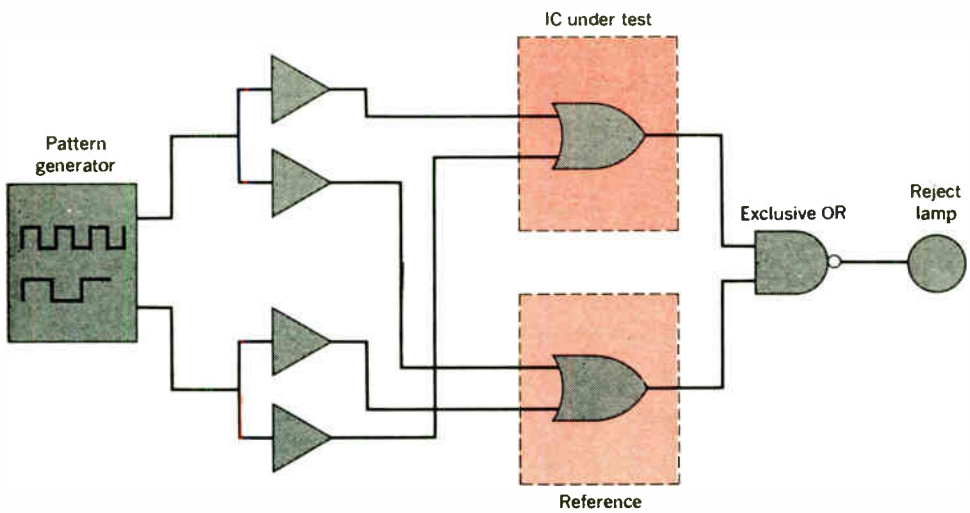


FIGURE 2. Block diagram of comparison-type functional tester.

Instructions to apply stimuli to the IC are buffered, converted into analog voltages, and delivered to the pins at the test sockets of multiplexed test stations (or to wafer probers), which are time-shared under computer control. The output functions of the IC are converted into digital form, buffered, and returned to the control unit and computer for processing. The operator exercises overall control of the system by teletypewriter keyboard commands.

Systems of the type shown in Fig. 1 typically cost from $50 000 to well over $100 000 and find greatest application among producers and high-volume users of ICs. At the other end of the scale is the comparison-type tester, shown in block-diagram form in Fig. 2. Here a binary or random pattern is applied to the device under test and at the same time to a reference unit having the same truth table as the unknown. The outputs are compared and a reject lamp lights when they differ. Instruments using this technique are relatively inexpensive and are preferred for incoming inspection of ICs in low or moderate volume.

In some applications, especially those in which in-

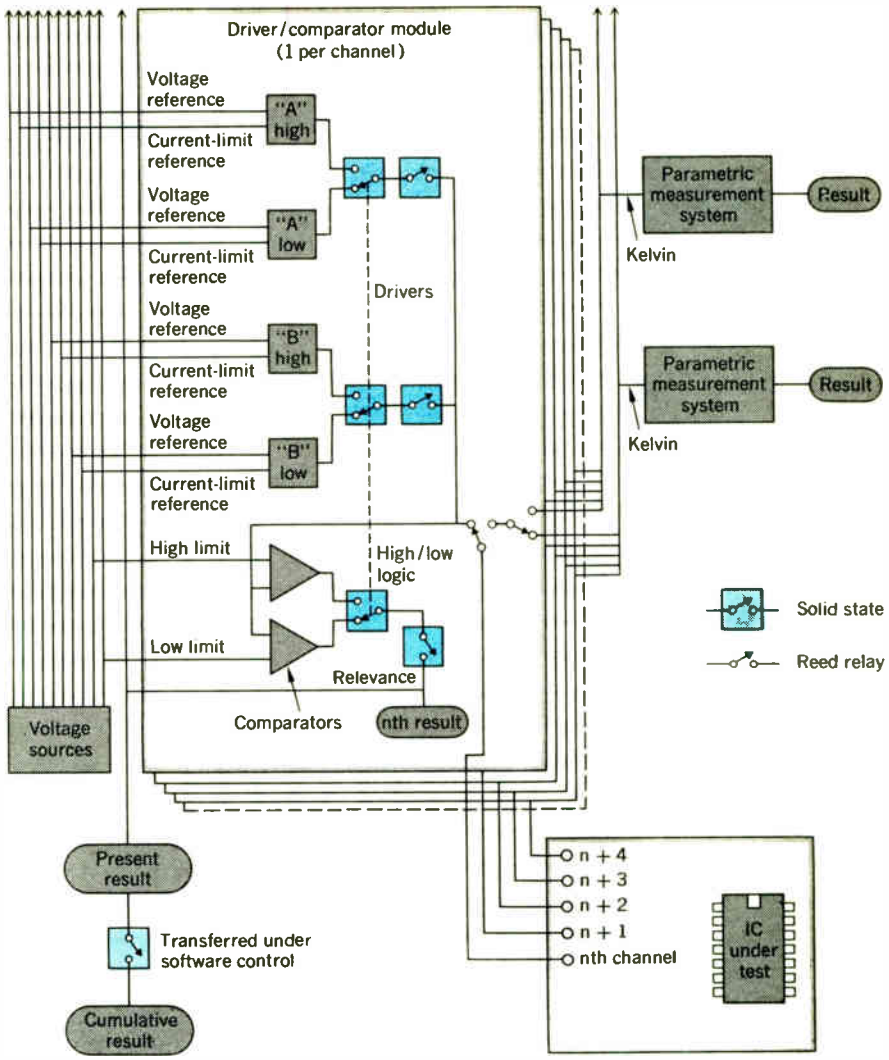


FIGURE 3. Driver/comparator switching, as employed in the Teradyne J283 computer-operated IC test system.

coming inspection is included, the ideal system or instrument would be versatile enough to handle bipolar and metal oxide semiconductor (MOS) devices, both digital and linear. Although such "universal" systems have been attempted, they are usually hybridizations of individual special-purpose test systems, based on the sharing of certain common elements (e.g., the controlling computer).

For all practical purposes, however, the testing of bipolar, MOS, and linear ICs can be considered three distinct subjects. The technologies are different, and the three types of devices represent three subindustries, each of which can afford to optimize its production processes.

Defining terms

After some years of semantic chaos, the lexicon of IC testing is beginning to stabilize. The principal branches of the tree are as follows:

Functional testing, which checks the truth table (or a subset of it) of a digital IC by applying a sequence of input words at nominal voltage levels and checking the corresponding output words. Functional testing usually involves a large number of tests and is therefore performed at the highest machine speed, at the expense of accuracy.

Clock-rate testing, which refers to the functional testing of MOS digital circuits at their maximum and minimum repetition rates.

Parametric testing, which measures IC voltages and currents at high accuracy and a relatively low test rate. Direct-current parametric testing refers to tests in which the inputs are maintained until the outputs reach a stable state. Pulse parametric (or dynamic) testing refers to tests of the time-related properties of an IC.

To summarize, the two chief classes of digital-IC testing are functional (high rate, low accuracy) and parametric (low rate, high accuracy). Note that the recording of a parametric value is *not* essential to parametric testing.

The testing of digital bipolar ICs

The digital bipolar class represents by far the largest segment of IC production today, and testing techniques are somewhat more standardized than they are for other device families. The usual pattern is: functional testing to find catastrophic failures caused by improper packaging, bonding, metalization, photolithography, die mounting, etc.; and dc and pulse parametric testing to uncover failures due to surface or oxide defects, such as channeling, pinholes, etc. Although virtually all ICs are tested func-

tionally and for dc characteristics, pulse parametric or dynamic testing is performed chiefly on fast transistor–transistor logic (TTL) or emitter-coupled logic (ECL) devices.

It is important to note that these three types of tests—functional, dc parametric, and pulse parametric—are related to distinctly different properties of an IC, and that a test sequence of one type only, no matter how thorough, cannot provide adequate device characterization.

Functional testing. A digital IC responds to a combination of high and low inputs (1's and 0's) by producing a certain combination of high and low outputs. Functional testing ensures that the combinations are as they should be for the logic in question. For combinatorial devices in which there are relatively few inputs, one can achieve thorough functional testing through the brute-force approach of exercising all input combinations. This approach breaks down, however, when the IC under test contains sequential logic, where the outputs are a function not only of the input combination but also of the order in which the various inputs are exercised.

The presence of sequential logic raises the number of possible input combinations and sequences far beyond the practical reach of even the fastest testers, and the testing problem then becomes one of choosing the best of the available compromises.

One compromise approach is based on the application of random patterns to the inputs and the statistical probability that these will test the device adequately. The shortcomings of this approach are that (1) the chances of testing for every possible failure mode are extremely remote, (2) a random pattern makes no allowance for time delays that flip-flops or other sequential devices may require at various points in their operation, and (3) the random pattern does not take into account the necessity for "initializing" certain ICs—that is, setting them to some known state before testing can begin.

Alternatively, one can algorithmically generate test patterns designed to detect all the failure modes intrinsic to the logic at hand. Software can be developed that will iteratively apply patterns, verify that given failure modes are or are not detected, and modify the patterns accordingly. This is a complex process and one on which much effort is being spent. Commercial pattern-generation services have sprung up in recent years to satisfy the growing demand for solutions to the testing problems associated with large-scale integration.

In the never-ending search for right combinations of 1's and 0's, it is all too easy to overlook the fact that an IC under test sees not 1's and 0's, but fast transitions of voltage or current. These transitions have to be fast enough to simulate the inputs the device will encounter in its end use and to represent decisive changes of state (i.e., a transition should not be so slow as to linger in the turn-on region of a device), but they should not be so fast as to produce unacceptable overshoot, ringing, or crosstalk, which can result in double-clocking of devices, channel interference problems, etc. An oscilloscope connected to the test points of a wafer prober will speak volumes about an IC test system's ability to test ICs reliably. Unless one can take a clean test signal for granted, he can never be confident of his test results, no matter how elegant the test patterns.

The functional-testing end of a computer-operated IC test system is diagrammed in Fig. 3. In this system (Teradyne's J283 "SLOT" system), each pin of the IC under test is connected to a module that contains two pairs of programmable "drivers" and a pair of comparators. The drivers are actually fast solid-state switches that gate power from buffered digital-to-analog (D/A) voltage sources. The voltage levels from these sources are assigned by computer control, and it is the function of the drivers to switch these voltages into the circuit quickly while preserving waveform integrity. All four drivers have programmable current limiting to prevent device damage.

The two comparators receive programmable reference voltages from the buffered source for use in determining whether IC output levels are above or below specified limits. Note that during functional testing each pin is always connected to both the driver and detector sections of the module, and that only a software command is needed to change a given channel from an input to an output or vice versa, an important consideration in the testing of certain ICs having pins that serve both functions. This arrangement also makes it possible for the system to apply a programmable load to an output pin during testing.

The output of either comparator is observed or not, depending on the presence or absence of a "relevance" software command. Thus, where one wishes to exercise an IC but ignore the logic outputs (as, for example, when preconditioning an IC), the comparator output is simply made nonrelevant. Where it is relevant, the system may be programmed to look for failures in any of three ways. It can look at each pin ("*n*th result"), sending pass–fail information back to the computer. Usually, however, it is not necessary to isolate failures down to pins and it is sufficient to know that some pin failed at a given point in the test sequence. Thus all the *n*th-result indications can be logically ORed to give a "present result." When long test patterns are applied even this method (which requires communication with the computer at each step) is impractical. In such cases the "present result" indication is automatically strobed into the "cumulative result" memory after each logic sentence. The cumulative result tells the operator that the IC failed somewhere along the line, which is very often the only information that is of interest.

Clock-rate testing. MOS clock-rate testing is analogous to functional testing, with one important difference: In bipolar functional testing, the test speed is very slow compared with the maximum speed at which the IC will operate, and thus is not a consideration. Clock-rate MOS testing, on the other hand, is conducted near the maximum frequency of the device, which, for today's faster devices, is in the region around 5 MHz, with 10 MHz over the not-too-distant horizon. At the other end of the spectrum, measurement of the "stay-alive" time (or minimum operating frequency) of the device may require a test frequency as low as 1 Hz.

Not only must the MOS test system be able to supply high-frequency test signals, it also must supply several sets of them (phases), each precisely settable with respect to the others. The number of phases needed depends on the types of devices to be tested; common requirements are for two or four phases with a phase resolution of 1 ns or better. The ability to manipulate phases with respect to

one another adds another dimension to the use of test patterns in LSI testing; an alternative to the use of many long patterns may be the application of a few worst-case patterns under varying phase relationships.

Great demands are placed on the drivers of an MOS test system. They must be able to swing 30 volts, at a slope of 1 ns/V or better, with minimum overshoot, ringing, or crosstalk, through cables to automatic handlers or wafer probers. Satisfactory performance results once one recognizes the practical necessity of such cables between drivers and the device under test and designs the test system accordingly, using impedance-matching techniques to minimize the effects of cable capacitance.

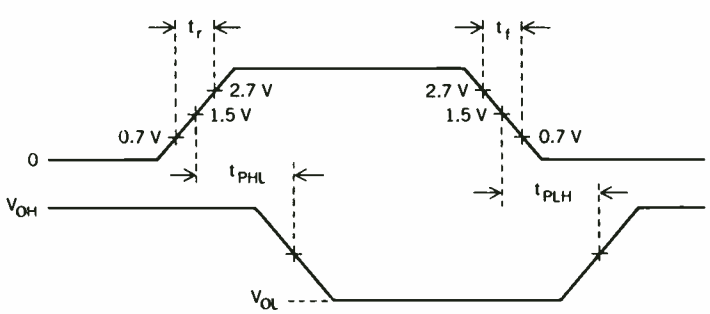
Parametric testing. Functional testing, even when exhaustively complete, cannot be relied on to determine whether an IC will operate in its end use. The test system cannot simulate all of the possible circuits in which a device may be used, and it is therefore necessary to measure certain parameters and to compare them against specified limits. These measurements will define the fanout capabilities of the device, as well as leakage current, power dissipation, etc. Usually a few parameters are measured for each of a number of input conditions.

The technique for making dc parametric tests is that of forcing a voltage or current at an input and comparing the resulting output current or voltage against a limit. The test result can be taken as a simple go/no-go indication, or A/D conversion techniques can be applied to record the actual value of the parameter in question. One such technique is a software-directed sequential approximation in which a series of go/no-go comparisons is made, the reference converging on the unknown.

Since the emphasis in parametric testing is on accuracy, precautions are taken in equipment design to eliminate stray capacitance and spurious ground currents. Kelvin connections are generally used, in conjunction with driven guard shields, to minimize cable charging currents that could introduce time-constant delays in circuit stabilization.

Because a parametric test generally takes much longer than a functional test, the interplay between the two types of test directly affects productivity. The programmer of a computer-operated system has several options available to him: He can run all functional tests first, in order to screen out catastrophic rejects before parametric testing; or he may make certain critical parametric tests first; or he may functionally test, branching into a parametric sequence upon failure.

FIGURE 4. Dynamic properties of a 5400-series TTL gate. Pulse parametric system measures rise and fall times and propagation delays t_{PHL} and t_{PLH}.



Whereas functional testing usually involves voltage swings of 30 volts or less, dc parametric test systems typically can force 100 volts or more. In systems having both functional and parametric test sections, break-before-make switching from one to the other is required so that the power available for parametric testing cannot inadvertently damage the functional-test drivers and comparators.

One of the most interesting and significant recent developments in IC testing has been the growing emphasis placed on pulse parametric, or dynamic, testing. Several factors lie behind this trend. First, speed margins represent the essential differences (and therefore the price premiums) between one device type and another. Second, these differences in operating speed cannot be verified by dc and functional testing. Third, equipment that can reliably measure dynamic performance on a production-line basis has become available only fairly recently.

Pulse parametric testing refers to a limited number of time-interval measurements—principally those of propagation delay, rise time, and fall time. For the fastest digital devices, these intervals are so short as to challenge the state of the measurement art. A test system handling ECL and fast TTL logic must be able to measure a propagation delay of a nanosecond repeatedly and with a precision of 10 picoseconds.

Some of the key issues in pulse parametric testing have to do with the way parameters are defined and specified. Rise time, for example, is often defined as the time it takes a voltage to rise from 10 to 90 percent of its maximum value, but, given a pulse with any overshoot or ringing, the maximum value and therefore the rise-time boundaries are uncertain. A much more rigorous definition would prescribe actual voltage levels as the boundaries for rise and fall times and propagation delay.

The parameters involved in a typical pulse parametric test are defined in Fig. 4, which illustrates the dynamic characteristics of a typical TTL gate. The rise and fall times of the input pulse, t_r and t_f, are defined in terms of actual voltage levels, not percentages; t_{PHL} and t_{PLH} are the propagation times from high to low and from low to high levels, respectively, and both are usually specified and tested.

Note that the accuracy with which one can define t_{PHL} and t_{PLH} depends on the accuracy with which the 1.5-volt thresholds are known, and this in turn is a function of the slope of the voltage transitions (a slow transition rate amplifies any threshold error). Input transition rate should therefore be specified, along with pulse amplitudes and durations.

Once the characteristics of the input pulse have been specified, the problem becomes one of ensuring that these characteristics are achieved, not at the output of the pulse generator but at the test socket. In a self-calibrating system each test pulse is first measured at the test socket and the pulse generator is automatically adjusted to produce the desired characteristics at the socket.

Early dynamic measurements on ICs were made by sampling techniques similar to these well established in high-frequency laboratory measurements. More recently, the "real-time" or "single-shot" technique, in which a single time interval is measured in terms of the amount of charge absorbed by a reference capacitor during that time, has achieved widespread acceptance and appears now to predominate.

To be useful on a production line, a dynamic test system must be compatible with automatic handling equipment and probers, and here the normal problems of preserving clean test waveforms are multiplied a thousandfold because the pulses involved are extremely fast. Another practical requirement is that dynamic tests be performed along with functional and dc tests, in a single insertion of the IC in the test socket.

Despite many technological obstacles, dynamic testing has now come of age. At least one computer maker has turned to 100 percent dynamic testing of all ICs in incoming inspection, in an effort to find dynamically defective devices before they can be loaded on circuit boards. The price of discovering defective devices after they already have been assembled in circuits is simply too great.

The testing of semiconductor memories

Semiconductor memories are functionally and parametrically tested like other ICs, but the fact that they have fixed outputs somewhat simplifies the problem. Various test patterns are used to ensure that the reading or the writing of a bit of memory will not affect the logic in adjacent cells. One commonly used pattern is a checkerboard of 1's and 0's. Another is the floating of a 1 or a 0 from cell to cell while the adjacent cells are maintained in the opposite state. There are many variations on these themes.

Where only a limited number of memory types are to be tested one may turn to a special-purpose memory exerciser such as that shown in Fig. 5. In this instrument, the Macrodata MD-100, the desired patterns are stored on read-only memories and the device signal and timing information is programmed on plug-in "personality cards."

The larger, computer-operated IC test system, on the other hand, brings the many advantages of computer control to memory testing. Most important of these is the ability to handle any semiconductor memory without the need for additional equipment.

The generation of test patterns for memories is simple in concept, but with even moderate-size memories the pattern length is long enough to be a consideration. Floating 1's and 0's through a 256-bit memory requires more than 120 000 functional tests, which are too many for economical storage in core or for fast disk-to-core transfer. One answer to this problem is on-line pattern generation, which allows a test system to produce patterns as it tests, using only about 1200 words of minicomputer memory.

The testing of linear ICs

Until a year or so ago, "linear IC testing" meant, for all practical purposes, the testing of operational amplifiers, which in turn meant the measurement of the offset voltage, offset current, open-loop gain, power capability, common-mode rejection, and a few other dc characteristics of these devices. More recently, however, several other types of linear IC have begun to appear in volume, many of them for consumer-electronics applications. Now linear IC testing includes the testing of circuits as diverse as voltage regulators and television chroma demodulators. Thus there is little that can be said of linear testing in general, beyond some observations on the types of test equipment commonly used.

For production testing, the advantages of computer control are overwhelming. There is, in fact, no real alternative; the number of parameters that may have to be measured at one time or another is well beyond the reach of a hardware-only system. Even with computer-operated systems, the variations in test requirements are so great that much of the test programming must take the form of "performance boards," which are special-purpose plug-ins incorporating the output loads and circuits needed to test a given family of devices.

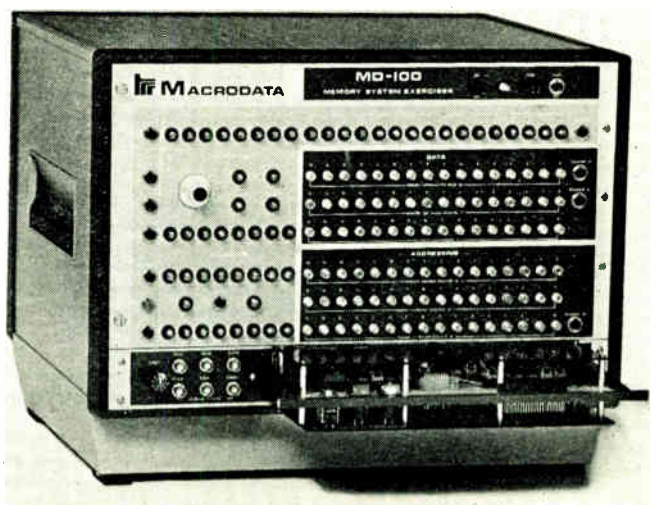
An interesting sidelight is that linear ICs are much more thoroughly tested than their nonmonolithic predecessors. Consider, for example, the list of tests performed by an automatic test system on a stereo demodulator in IC form:

1. Power consumption.
2. Left and right output level.
3. Audio mute on and off thresholds.
4. Audio mute attenuation.
5. Audio mute thump.
6. Channel balance.
7. Rejection tests for 19- and 38-kHz components.
8. Stereo separation.
9. Monaural distortion.
10. Stereo distortion.
11. Lamp on and off threshold.
12. Stereo muting on and off threshold.

This entire complement of tests, plus an automatic pilot-subcarrier phase adjustment, can be made in well under a second by a computer-operated test system. It is doubtful that any discrete-component stereo demodulator was ever as thoroughly tested in production, a good deal of manual trimming and tweeking notwithstanding.

Although the more advanced consumer circuits must be tested by large computer-operated systems, simpler devices such as operational amplifiers, voltage regulators, and comparators can be tested in incoming inspection by small bench-type instruments, typically programmed by plug-in circuit boards. One such instrument, the General Radio Type 1730 Linear Circuit Tester, is shown in Fig.

FIGURE 5. A noncomputer-operated tester for semiconductor memories (the Macrodata MD-100).



6. This instrument is programmed by a plug-in card containing a bank of 40 resettable slide switches.

General considerations in the evaluation of IC test equipment

The first requirement of any testing, productivity considerations notwithstanding, is that the tests be valid. In IC testing this means (1) identifying those parameters that adequately define the "goodness" or "badness" of a device, (2) ensuring that the conditions of measurement relate meaningfully to the conditions of use, and (3) achieving the desired accuracy of measurement.

Specifying the accuracy of IC measurements is an especially precarious business because the test environment is unpredictable—and often hostile. Extremely high accuracy and precision are rarely primary considerations, since the required levels are usually well within the state of the art. Instead, the emphasis is on repeatability, with periodic verification of accuracy. The word "calibration" is anathema to most test-system users, because it implies downtime. When a system is found to be outside tolerance one approach is isolation and quick replacement of the guilty component.

Although accuracy and precision—in the traditional instrument sense—are not primary issues, the "analog" performance of an IC test system most certainly is. A "clean" transition, without overshoot or ringing, is essential if fast devices are to be exercised without ambiguity. The signal should be clean at the device under test, which in production testing usually means at a prober or handler separated from the signal source by a length of cable.

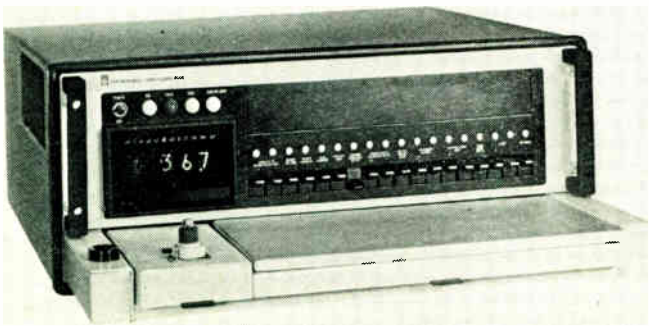
Productivity

Since economics are such a dominant factor in semiconductor production, one tends to think not of the cost of a tester but of the cost of *testing*. Thus productivity is a much-used term and a much-sought-after characteristic of test equipment. Productivity in testing is a function of many variables, including the following:

1. *Test speed.* Although speed is the most obvious factor, it is rarely the most important. Even the slowest test system is so fast that the real productivity limitations arise elsewhere.

2. *Multiplex capability.* Many test systems provide for some degree of multiplexing, or the simultaneous use of two or more test stations with a single test system.

FIGURE 6. An incoming-inspection instrument for linear ICs (the General Radio 1730 Linear Circuit Tester).



The number of stations and the level of independence of each directly affect productivity. A four-station system, for example, may or may not be able to distribute its services among three probe stations and one final test station, or among four stations testing four different devices, or among three classification stations and one data-logging station.

3. *Handling speed.* If devices are manually inserted into a test socket of a single-station system, the operator's handling rate will almost certainly be the major factor limiting throughput. It is the great imbalance between testing and handling speeds, in fact, that provides the rationale for multiplexing.

The use of automatic handling equipment speeds up the mechanical end of the process to the point where several thousand devices an hour can be tested at a single station, assuming that the test sequence per device is fairly simple.

The fastest mechanical handler cannot begin to match the test system's speed for simple devices (e.g., well under 100 ms to test a gate). Where automatic handlers are to be multiplexed, however, test time and handling converge, and the relation between the two must be noted in setting up the test installation and program.

4. *Programming and debugging.* A computer-operated test system is normally supplied with an executive program that contains the formatting and subroutines needed for a given type of application (end-of-life testing, classification, etc.). The user must write and enter his own test specifications (bias conditions, limits, bin criterions, etc.), and then he must verify and correct ("debug") his program as necessary. The debugging procedure can easily consume great amounts of time (an hour a day is not at all unusual), so a test system must be evaluated in the light of its ability to time-share programming and debugging with normal testing and the availability of debugging aids (panel indicators, check-sum verification of software, etc.).

5. *Downtime.* Downtime is the total time during which a test system is not available for normal testing. In any calculation of productivity, downtime obviously comes right "off the top," all other factors coming into play only when the system is operative. It is a difficult parameter to quantify when making a purchasing decision, but it is so important that an effort must be made. Consider the difference in productivity between a test system regularly experiencing 20 percent downtime (a not unusual figure) and one with 1 percent downtime (also a not unusual figure).

6. *Retest load.* If a system is not testing devices properly, the lots rejected by quality assurance (QA) or incoming inspection will usually find their way back to be tested again. A system that is out of calibration will therefore directly penalize system productivity.

7. *Test-plan efficiency.* In most cases, the number of tests required to test an IC adequately is not subject to rigid definition. One can determine easily enough the number of truth-table combinations for a simple IC, but where this number becomes overwhelmingly large, attention shifts to the number required for adequate (rather than exhaustive) testing, leaving room for wide system-to-system variations. Judgment dictates the proper extent of dc parametric or linear testing per device, but system flexibility converts this judgment into productivity. If the testing objective is usually to characterize an IC as

"good" or "bad," then it is not enough to reduce the total number of tests needed from, say, 50 to 40 per device. Why bother making 40 tests on a bad device? Why not arrange the order of tests so that bad devices are disqualified as soon as possible—perhaps even after one or two tests? This suggests placing the most critical tests first in the sequence (e.g., testing for catastrophic failures, then functionally testing, then parametrically); but there is more to it than that. If a given test, no matter how critical, is almost always passed, then it is a waste of time to put it at the head of the sequence. The possibilities are seen to be endless in theory, but in practice they are limited by the flexibility of the test-system hardware and software.

8. *Data logging.* Although the concept of productivity in semiconductor testing usually refers to device throughput, it is also validly applied to the rate of generation of test data (summary sheets, distribution analyses, end-of-life data, etc.). The speed of the data-logging medium (teleprinter, line printer, etc.) is important, though less so than one might think: Much data logging is quickly stored on magnetic tape for later, off-line printout.

The foregoing considerations apply chiefly to large test systems in production applications and to a lesser degree to IC test instruments used in inspection or engineering. A production test system, remember, is as much a part of the manufacturing process as a diffusion furnace. Since no one is really happy with a manufacturing facility operating below capacity, the ideal condition is one in which the test equipment, along with everything else, is being pushed to its limit. In most inspection operations, on the other hand, test equipment is purchased not because a plant would cease to function without it but in order to reduce costs of equipment rework and service. An inspection instrument can be idle half the time and still easily earn its keep. This is not to suggest that those who manage inspection facilities are not concerned with productivity; they are, but they define productivity in broader terms, including the total time spent on the end product.

Versatility and obsolescence

Any realistic appraisal of IC technology would have to conclude that, far from being a mature field, it has yet to approach full development. The 300 million ICs produced in 1970 include hardly any for consumer and automotive electronics, two areas of major potential. (It has been estimated that the automotive industry alone could consume 200 million ICs annually by 1974.) New applications breed basic device changes. So does the current heavy expenditure in semiconductor R&D. Variations on the MOS theme seem to occur weekly. All of this seems to threaten all existing IC test equipment with early obsolescence.

Yet most of the several hundred production test systems shipped during 1967 and 1968 are still hard at work for their owners. How can a system designed for 1967 IC technology handle 1971 devices? The answer lies in the fact that the technological ground rules for IC testing have been fairly well established for years. New types of devices require new testing technology, to be sure, but a gate today is tested very much the same way a gate was tested several years ago. The advances have come in the *number* of gates that can be tested per unit time and in overall system performance parameters. Yesterday's IC test systems are helped greatly in their battle against obsolescence by the fact that many of them are computer-controlled. Give a computerized system a new software package and it is effectively a new system. (For this reason, if for no other, the computer-operated system represents more actual value than a hard-wired or tape-programmed system.) For all its newly acquired intelligence, however, the older system is not likely to be as productive as newer systems, so its versatility can take it only so far.

System compatibility

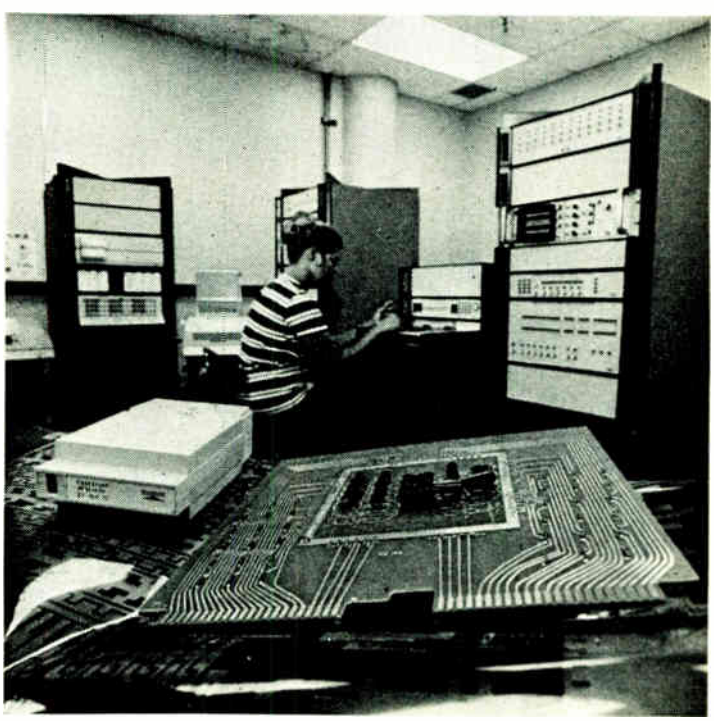
Although some bench-top IC testers operate "barefoot" in incoming inspection facilities, IC testing is by and large a systems operation, and the typical IC test system finds itself surrounded by wafer probers, ovens, automatic handlers, scanners, recorders, and other handling and processing paraphernalia. Thus the designer of an IC test system must take into account the fact that a wafer prober will introduce probe capacitance, that long leads will be required to connect to automatic handlers, that the user may want to add auxiliary instruments to the setup, etc.

It is reasonable to expect the maker, rather than the buyer, of an IC test system to assume most of the interface worries. The vagaries of automatic handlers and probers are well known to any established producer of IC test equipment, and it makes little sense for a system buyer to learn them the hard way.

Easy programming

The phrase "easy programming" has been applied to every commercial test system, to the point of meaning-

FIGURE 7. Magnetic-tape cartridge carries test program and "device board" houses output loads and other specialized test hardware for use with Teradyne J277 MOS Test System, in background.



lessness. In fact, writing the program (once one decides what he wants to test) is a fairly simple matter with any system. A language and some syntax must be learned, but the learning generally requires only a day or two, and sometimes much less. One is ill-advised to select a test system because it is easier to write programs for it than for another system. In fact, the easier language may reflect a more limited testing capability. The most effective utilization of a test system normally results when a good technician, having learned all the nuances of an *extensive* system language, calls into play all the testing power at his command.

The concept of easy programming is more valid, however, when applied to the type (rather than the size) of a programming language. Languages are sometimes characterized as "high level" or "low level." A high-level language is designed exclusively for the testing job at hand, whereas a low-level language is for general-purpose use. A high-level language geared to IC-testing terminology obviously will be easier to program in that application than a low-level language. Consider, for example, the programming of a pulse parametric test on Teradyne's J277 MOS Test System; see Fig. 7. To set up the timing of clocks and output strobe the operator might write:

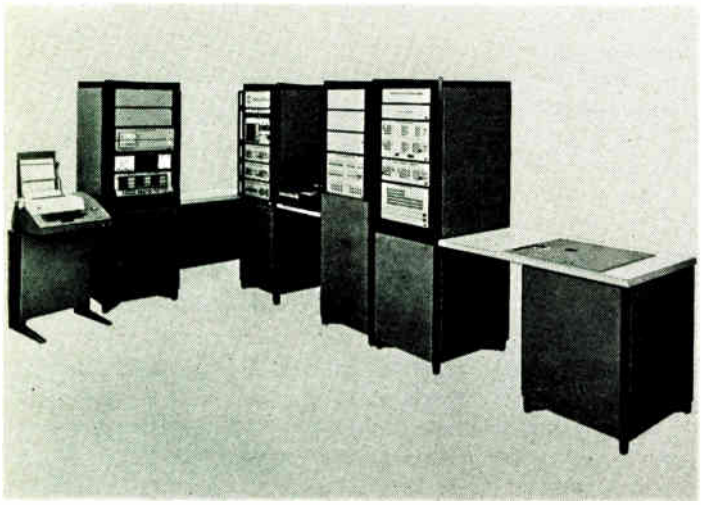
```
CLOCK 1     100 NS FROM T0
GAP 1       70 NS FROM CLOCK 1
CLOCK 2     210 NS FROM GAP 1
STROBE 1    300 NS FROM T0
```

(In this example a gap is defined to indicate that the clocks do not overlap.) The operator commands the system to measure the leakage current on pin 2, rejecting anything more negative than −500 mA at −18 volts, as follows:

```
MCON PIN 2
MEASURE −500 MA AT −18V; RNEG
```

Not all test-system languages are as close to plain English as the foregoing, but many use mnemonics that are easily mastered.

FIGURE 8. Two computer-operated test systems. (Top) Teradyne J283/Si57 system for functional, dc, and pulse parametric testing. (Bottom) Datatron 4400 for functional and dc parametric testing.



Sizing the equipment to the job

Most of the points discussed so far relate chiefly to the use of IC test systems. The world of IC testing, however, embraces bench-top IC testers as well as large, production-oriented systems. In some ways the smaller instruments are the equal of their massive counterparts. A bench-top tester may well be able to perform functional tests on a given IC as fast as any large system can. Small testers are in fact sometimes interfaced to systems, in order to screen out functional failures as fast as possible, before proceeding to parametric testing.

At the low end of the price spectrum of IC test equipment is the manually programmed tester costing a few hundred dollars. It can functionally check a simple IC, and that is about all. It is useless for the testing of ICs in any volume.

A fairly large selection of equipment is available in the $3000–$8000 range. Instruments in this class are typically comparison testers, programmed by plug-in printed-circuit cards, with a binary generator handling the functional-testing requirements at very respectable speeds. Instruments of this type are not to be taken lightly. Interfaced to an automatic handler, a $5000 IC tester can throw several thousand ICs an hour into "good" and "bad" bins, on the basis of both functional and parametric test results.

In determining whether or not a bench-top instrument will satisfy one's testing needs, the best approach is probably to examine the things that such an instrument *cannot* do, then determine whether any of these are crucial in a given application. Specifically, and with very few exceptions, the bench-top tester *cannot* generate test data; match the multiplexing ability of a large system; provide for dynamic testing; handle as many different types of IC as a computerized system can; perform thorough functional testing of devices having sequential logic; make highly accurate and precise parametric tests; or make burn-in or life tests.

These limitations rule out the bench-top tester for production and QA testing and incoming inspection where test data are required. Once it is determined that the bench-top tester is inadequate for the job at hand, the next step up the performance scale is the $50 000-and-up computer-operated test system. (An exception is

the noncomputer-operated memory tester at $15 000–$20 000.) A fully equipped, computer-operated dc test system for bipolar devices can be purchased for well under $100 000. Addition of a dynamic testing capability raises the price to the $100 000 to $150 000 range. Computerized systems for the clock-rate testing of MOS begin at about $100 000. Two representative computer-operated test systems are shown in Fig. 8.

Support costs for an IC test system—the costs of maintenance, calibration, programming, special test fixturing, etc.—vary widely, in many cases exceeding the initial system cost within the first year of operation. There are many ways of minimizing such support costs, and most of them bear on the selection of the test system. The desire to reduce programming, debugging, and service costs explains the banks of indicator lamps on the front panels of some test systems. These lamps are of no use during normal operation but are indispensable in setting up and checking the system.

Conclusion

The field of IC testing has been a showcase for computer control and automation. Much of the system expertise that has been developed is applicable to electronics production in general. This implies great opportunities for increased productivity in our industry, which is, curiously, one of the least automated of all.

Appendix: Manufacturers of IC test equipment

IC test equipment is commercially produced by about two dozen firms. New developments come so fast in this field that it would be neither helpful to the reader nor fair to the manufacturers to attempt to list equipment specifications. Names and addresses of the major manufacturers are given below for the benefit of those interested in obtaining information on the characteristics and application of IC test equipment.

Adar Associates, Inc., 85 Bolton St., Cambridge, Mass. 02140 (computer-operated bipolar and MOS test systems)

Alma Corp., 570 Del Rey Ave., Sunnyvale, Calif. 94086 (digital IC test instruments)

Datatron, Inc., 1562 Reynolds Ave., Santa Ana, Calif. 92711 (computer-operated bipolar and MOS test systems)

E-H Research Laboratories, Inc., 515 11th St., Oakland, Calif. 94604 (computer-operated bipolar and MOS test systems)

Fairchild Systems Technology, 974 E. Arques Ave., Sunnyvale, Calif. 94086 (computer-operated MOS, bipolar, and linear test systems; linear IC test instruments)

General Radio Co., 300 Baker Ave., Concord, Mass. 07142 (linear test instruments)

LSI Testing, Inc., 2280 S. Main St., Salt Lake City, Utah 84115 (computer-operated bipolar and MOS test systems)

Macrodata, Inc., 20440 Corisco St., Chatsworth, Calif. 91311 (computer-operated MOS and bipolar test systems; memory test instruments)

Microdyne Instruments, Inc., 203 Middlesex Turnpike, Burlington, Mass. 01803 (digital and linear test instruments)

Optimized Devices, 220 Marble Ave., Pleasantville, N.Y. 10570 (linear test systems)

Redcor Corp., 21200 Victory Blvd., Woodland Hills, Calif. 91364 (computer-operated MOS test systems)

Tektronix, Inc., P.O. Box 500, Beaverton, Oreg. 97005 (computer-operated MOS and bipolar test systems)

Teradyne, Inc., 183 Essex St., Boston, Mass. 02111 (computer-operated MOS, bipolar, and linear test systems; digital test instruments)

Teradyne Dynamic Systems, 9551 Irondale Ave., Chatsworth, Calif. 91311 (computer-operated bipolar test systems)

Watkins-Johnson, Inc., 3333 Hillview Ave., Palo Alto, Calif. 94304 (computer-operated bipolar test systems)

Western Digital, 1612 S. Lynn St., Santa Ana, Calif. 92705 (computer-operated bipolar and MOS test systems)

Xintel Corp., 20931 Nordhoff St., Chatsworth, Calif. 91311 (computer-operated bipolar and MOS test systems)

BIBLIOGRAPHY

Attridge, W. A., "Caution: test op amps carefully," *Electron. Design*, Nov. 8, 1969.

"Automatic and manual integrated circuit test equipment," SETE 210/96, Project SETE, New York University, Aug. 1968.

Beck, R., "Functional-testing an IC memory," Application Rept. 105, Teradyne, Inc.

Bourne, R. B., Jr., "Fault detection in bipolar integrated circuit outputs," Application Rept. 113, Teradyne, Inc.

Boyle, A. J., "Testing MOS," *Electron. Engr.*, Oct. 1970.

Curran, L., "Meeting the MOS/LSI challenge: a special report on testers," *Electronics*, May 10, 1971.

Edelman, S., "Testing integrated circuits," *Electron. Engr.*, Sept 1970.

Egan, F., and Speer, R., eds., "IC testing—a special report," *Electron. Design*, Sept. 1, 1968.

Flynn, G., "Forum on op amps," *Electron. Products*, Dec. 1968.

Padwick, G. B,, "Dynamic IC testing made easy," *Electronics*, Sept. 30, 1968.

Salvador, J., "Today's dynamic IC tests won't work without meaningful specs," *Electronics*, Nov. 8, 1971.

Seaton, J., "Testing low input currents in operational amplifiers," Application Rept. 106, Teradyne, Inc.

Young, F. M., "Clock-rate testing in a realistic environment," *1971 WESCON Rec.*

Reprints of this article (No. X71-122) are available to readers. Please use the order form on page 8, which gives information and prices.

Frederick Van Veen (SM) received the B.A. degree from Boston College in 1951. After graduation he served as an officer in the U.S. Army Reserve, completing various courses in electronics at the Artillery School, Fort Bliss, Tex. He joined the General Radio Company in 1955; later he became publicity manager and editor of the General Radio Experimenter. In 1968 he joined Teradyne, Inc., where he is now director of corporate relations. He has written extensively on the subjects of electronic instrumentation and semiconductor testing. Since 1965 he has served as editor of the IEEE Transactions on Audio and Electroacoustics. He is also past national chairman of the IEEE Engineering Writing and Speech Group.

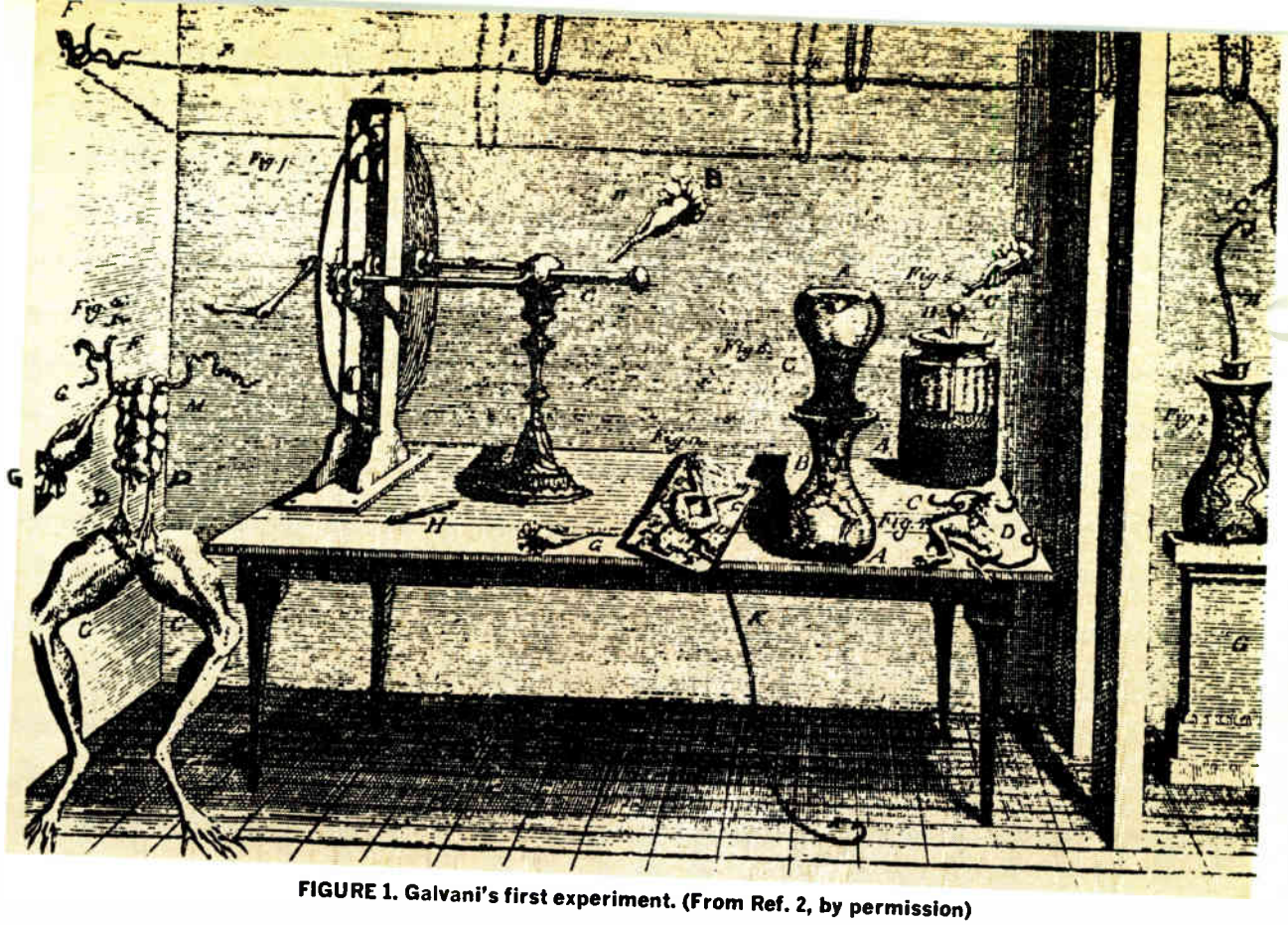


FIGURE 1. Galvani's first experiment. (From Ref. 2, by permission)

The discovery of bioelectricity and current electricity

The Galvani–Volta controversy

Although the history of static electricity goes back to the ancient Greeks, prior to the era of Galvani and Volta no source had been discovered that could deliver a continuous electric fluid—a term that implies both charge and current

L. A. Geddes, H. E. Hoff Baylor College of Medicine

Amid the tumultuous confusion and complexity of modern events, it is difficult to acquire the general historical perspective so important to the scientist and the engineer of pioneering nature. Yet a broad vision of the past is a most valuable asset to the workers and inventors of today. It forms a foundation for further progress and suggests roads to even higher goals than we have thus far reached. It is truly said that we of today stand on the shoulders of giants.

To help us use the past to inspire us to future achievement, the authors of this article give us a thrilling trip in what might be termed a "technological time machine." Hopefully we shall all benefit from the lessons set forth, from the examples of Galvani and Volta, of Nobili and Matteucci. We of the space age, with all the array of modern techniques at our command, clearly still can gain much by noting and emulating the ingenuity, inspirational qualities, and tenacity of purpose of these great pioneers.

Present-day investigators may also marvel at the insight of their predecessors who derived so much fundamental knowledge with the primitive and even erratic equipment at their disposal. It is truly remarkable that they wrested such great wisdom from such limited resources. If the story of their ability can continue to guide and encourage the scientist and the engineer, then the future promises ever-greater accomplishments for the service of mankind.

A. N. Goldsmith
Director Emeritus, IEEE

Although almost everyone has heard of Galvani and Volta—whose names have entered the English language in such words as galvanize and volt—surprisingly few have an accurate knowledge of the experiments they performed, the conclusions they drew therefrom, and the profound manner in which their studies influenced science within only a few years. There is no doubt that the chance observation by Galvani, the three experiments that he conducted, and the controversy that ensued with Volta were at the basis of the discovery of all bioelectric phenomenons as well as of current electricity, despite the fact that Galvani's explanation for his experiment was wrong and Volta was not altogether correct in the theory he presented for the operation of the battery he devised. Because of the importance of these events, the authors have sought to present the evidence to the reader by recounting the Galvani–Volta story in a simple way, using quotations drawn directly from Galvani's and Volta's reports. In addition, the phenomenons they discovered are analyzed in the light of present-day knowledge.

The critical events of this history took place in the decade just prior to 1800. Before this time there had accumulated a considerable inventory of facts relating to bioelectricity and electricity in general. The Egyptians and Greeks had known that certain fish could deliver substantial shocks to an organism in their aqueous environment.[1] It was also known that the application of generators of static electricity would cause muscles to twitch by stimulating the muscle or its nerve directly. Well before the dawn of Christianity, static electricity had been discovered by the Greeks, who produced it by rubbing resin (amber or, in Greek, elektron) with cat's fur. Later it was found that static electricity could be created by rubbing glass with silk. It was thought that different types of electric "fluid" were produced by these two methods; consequently the adjectives "resinous" and "vitreous" were applied. It was also common knowledge that a substantial quantity of electric fluid could be produced by frictional electric machines, the first of which appears to have been developed by Von Guericke about 1672. In 1747 Van Muschenbroek and Von Kleist independently discovered that electric fluid could be stored in a Leyden jar, the precursor of the capacitor.

About this time Franklin had become interested in electricity and he performed experiments that led to the "one-fluid theory," which stated that there was but one type of electricity and that the electrical effects produced by friction reflected the separation of electric fluid so that one body contained an excess, the other a deficit. (It is interesting to note that the "unit of electric fluid," the electron, was not discovered until more than a century later.) Franklin also showed that atmospheric electricity (lightning) and artificial electricity, derived from electric machines, were one and the same. By this time Coulomb had performed his experiments and enunciated his inverse-square law for the force of attraction or repulsion between charged bodies. Despite this array of discoveries, it is important to note that before the time of Galvani and Volta, there was no source that could deliver a continuous flow of electric fluid, a term that implies both charge and current.

Galvani's experiments

Against such a background of knowledge of the "electric fluid" and the many powerful demonstrations of its ability to activate muscles and nerves, it is readily understandable that biologists began to suspect that the "nervous fluid" or the "animal spirit" postulated by Galen to course in the hollow cavities of the nerves and mediate muscular contraction, and indeed all the nervous functions, was of an electrical nature. Galvani, an obstetrician and anatomist, was by no means the first to hold such a view, but his experimental search for evidence of the identity of the electric and nervous fluids provided the critical breakthrough. During the first of his experiments, conducted in his home, he noted that every time a spark was drawn from a nearby static-electricity machine while an assistant was touching the sciatic nerve trunk of a frog with the point of a scalpel, the leg muscles twitched. Green's translation[2] of Galvani's account follows:

"I dissected and prepared a frog [as in Fig. 1] and placed it on a table, on which was an electrical machine... widely removed from its conductor and separated by no brief interval [distance]. When by chance one of those who were assisting me gently touched the point of a scalpel to the medial crural nerves, *DD*, of this frog, immediately all the muscles of the limbs seemed to be so contracted that they appeared to have fallen into violent tonic convulsions. But another of the assistants, who was on hand when I did electrical experiments, seemed to observe that the same thing occurred whenever a spark was discharged from the conductor of the machine...."

After some additional experiments to authenticate the phenomenon, Galvani wrote:

"Aroused by the novelty of the circumstance, we resolved to test it in various ways, and to experiment, employing nevertheless the same scalpel, in order that, if possible, we might ascertain the causes of the unexpected difference; nor did this new labor prove vain; for we found that the whole thing was to be attributed to the different part of the scalpel by which we held it with our fingers: for since the scalpel had a bone handle, when the same handle was held by the hand, even though a spark was produced, no movements resulted, but they did ensue, if the fingers touched either the metallic blade or the iron nails securing the blade of the scalpel."

It is clear from this report that Galvani didn't demonstrate the existence of bioelectricity; stimulation of the crural (sciatic) nerve occurred by electrostatic induction in the circuit between the static-electricity machine, the frog preparation, the observer, and earth. Nonetheless, the experience prompted him to wonder if atmospheric electricity would produce a similar response. His next experiment was designed to test this hypothesis. The various translations of Galvani's second experiment agree in overall, but not fine, detail; Green's translation[2] states:

"Having discovered the effects of artificial electricity on muscular contractions which we have thus far explained, there was nothing we would sooner do than to investigate whether atmospheric electricity, as it is called, would afford the same phenomena, or not: whether, for example, by employing the same devices, the passage of lightning, as of sparks, would excite muscular contractions.

"Therefore we erected, in the fresh air, in a lofty part of the house, a long and suitable conductor, namely an iron wire, and insulated it [Fig. 2] and to it, when a storm arose in the sky, attached by their nerves either prepared frogs, or prepared legs of warm animals .... Also we attached another conductor, namely another iron wire, to the feet of the same, and this as long as possible, that it might extend

as far as the waters of the well indicated in the figure. Moreover, the thing went according to our desire, just as in artificial electricity; for as often as the lightning broke out, at the same moment of time all the muscles fell into violent and multiple contractions, so that, just as the splendor and flash of the lightning are wont, so the muscular motions and contractions of those animals preceded the thunders, and, as it were, warned of them; nay, indeed, so great was the concurrence of the phenomena that the contractions occurred both when no muscle conductor was also added, and when the nerve conductor was not insulated, nay it was even possible to observe them beyond hope and expectation when the conductor was placed on lower ground ... particularly if the lightnings either were very great, or burst from clouds nearer the place of experimentation, or if anyone held the iron wire *F* in his hands at the same time when the thunderbolts fell."

In the course of these experiments, Galvani noted another phenomenon, which, according to Green,[2] he reported as follows. This observation has become known as Galvani's second experiment.

"Wherefore, since I had sometimes seen prepared frogs placed on iron gratings which surrounded a certain hanging garden of my house, equipped also with bronze [some accounts refer to copper or brass] hooks in their spinal cord, fall into the customary contractions, not only when the sky was lightning but also sometimes when it was quiet and serene, I thought these contractions derived their origin from the changes which sometimes occur in atmospheric electricity. Hence, not without hope, I began diligently to investigate the effects of these changes in these muscular motions in various ways. Wherefore at different hours, and for many days, I inspected animals, appropriately adjusted therefor; but there was scarcely any motion in their muscles. Finally, weary with vain expectation I began to press the bronze hooks, whereby their spinal cords were fixed, against the iron gratings, to see whether by this kind of device they excited muscular contractions, and in various states of the atmosphere, and of electricity whatever variety and mutation they presented; not infrequently, indeed, I observed contractions, but bearing no relation to varied state of stmosphere or of electricity.

"Nevertheless, since I had not inspected these contractions except in the fresh air, for I had not yet experimented in other places, I was on the point of seeking such contractions from electricity of the atmosphere, which had crept into the animal and accumulated in him and gone out rapidly from him in contact of the hook with the iron grating; for it is easy in experimentation to be deceived, and to think one has seen and discovered what we desire to see and discover.

"But when I had transported the animal into a closed chamber and placed him on an iron surface, and had begun to press against it the hook fixed in his spinal cord, behold the same contractions and the same motions! Likewise continuously, I tried using other metals, in other places, other hours and days; and the same result; except that the contractions were different in accordance with the diversity of metals, namely more violent in some, and more sluggish in others. Then it continually occurred to me to employ for the same experiment other bodies, but those which transmit little or no electricity, glass for example, gum, resin, stone, wood, and those which are dry; nothing similar occurred; it was not possible to observe any muscular motions or contractions. Results of this sort both brought us no slight amazement and began to arouse some suspicion about inherent animal electricity itself. Moreover both were increased by the circuit of very thin nervous fluid which by chance we observed to be produced from the nerves to the muscles, when the phenomenon occurred, and which resembled the electric circuit which is discharged in the Leyden jar."

Figure 3 illustrates a variety of Galvani's experiments in which a bimetallic arc is used to connect the spinal marrow to the leg muscles. Although Galvani noted that the "contractions were different in accordance with the diversity of the metals," he reasoned that by connecting the nerve and the muscle by the metallic arc (armature) consisting of two dissimilar metals, he had discharged the animal electricity present in the muscle, thinking that the nerve and muscle were analogous to the inner and outer conductors of a Leyden jar. It was Galvani's failure to realize that a conducting arc of two dissimilar metals was essential for contractions that later led Volta to find a very different explanation for the phenomenon.

(It is of interest to note that a new English translation of Galvani's *De Viribus Electricitatis* is available, with an introduction by Cohen of Harvard and a revised bibliography by Fulton and Stanton of Yale.[3] See also Ref. 4.)

Volta's explanation

Galvani's investigations aroused a virtual furor of interest. Wherever frogs were to be found, scientists and laymen alike repeated his experiments with routine success. Dibner[4] reports that Galvani's second experiment was performed at social gatherings and Galvani's explanation for the muscular contractions (discharge of a nerve–muscle Leyden jar) was accepted without question—even by Volta, who had received a copy of Galvani's paper and verified the phenomenon. However, Volta very soon had second thoughts on the subject and found that the essential requirement for producing contractions was the presence of two dissimilar metals joined at one end, with their free ends applied to either the nerve or muscle. If a single metal was applied to the nerve and muscle, no contractions were obtained. Volta immediately described his findings in letters to various learned societies. (The experiments conducted by Galvani and Volta were reported in letters and scientific publications.[5,6]) Of the various translations published, that which appeared in 1793 in the *Transactions* of the Royal Society—of which Volta was a foreign member—is perhaps one of the more interesting. The report was the content of a letter from Volta to Tiberius Cavallo[7]:

"But if he [meaning Galvani] had but a little more varied the experiments, as I have done, says Mr. Volta, he would have seen that this double contact of the nerve and muscle, this imaginary circuit, is not always necessary. He would have found, as I have done, that we can excite the same convulsions and motion in the legs, and the other members of animals, by metallic touchings, either of 2 parts of a nerve only, or of 2 muscles, and even of different points of one simple muscle alone."

Following a discussion of various complementary experiments, the report continues:

"Yes it is a quite different sort of method of electric fluid, of which we ought rather to say we disturb the equilibrium, than restore it, in that which flows from one

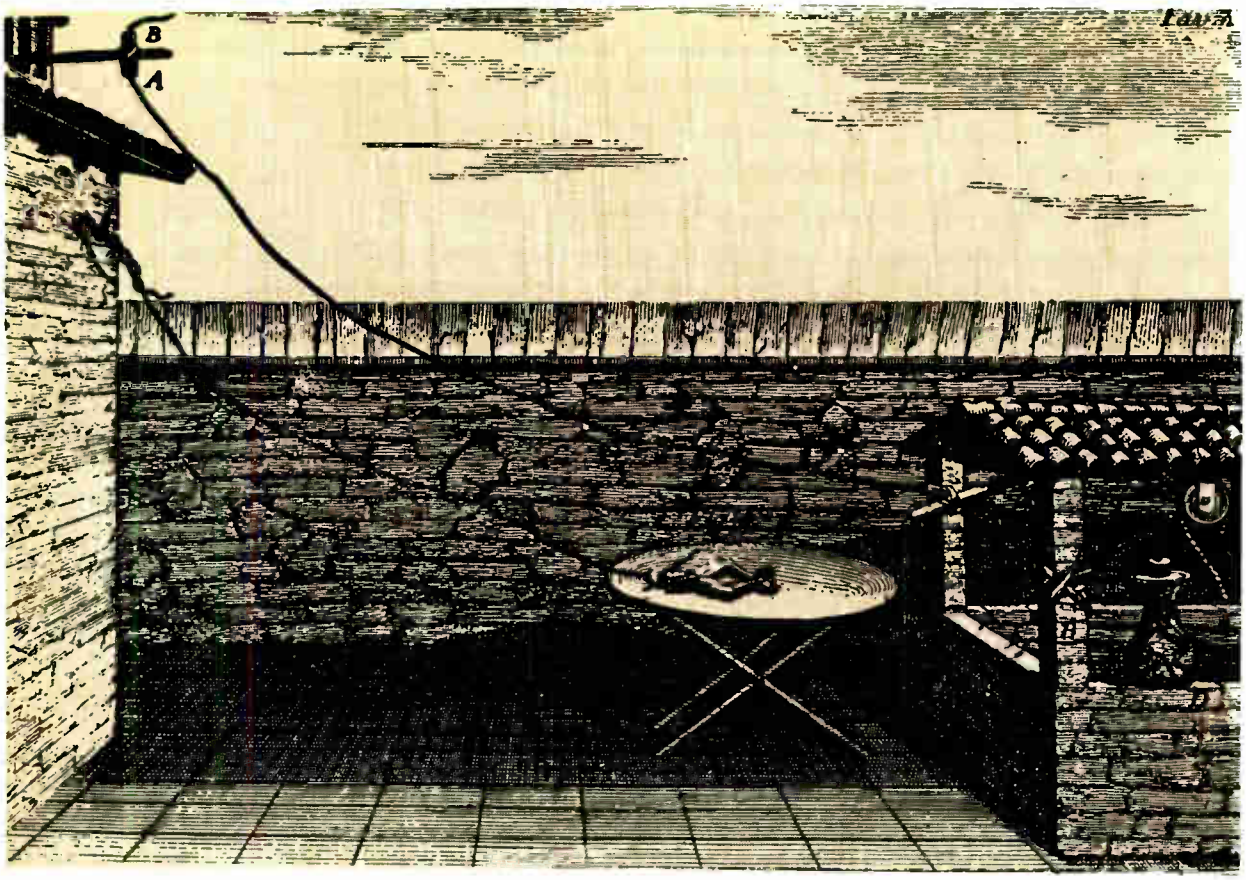
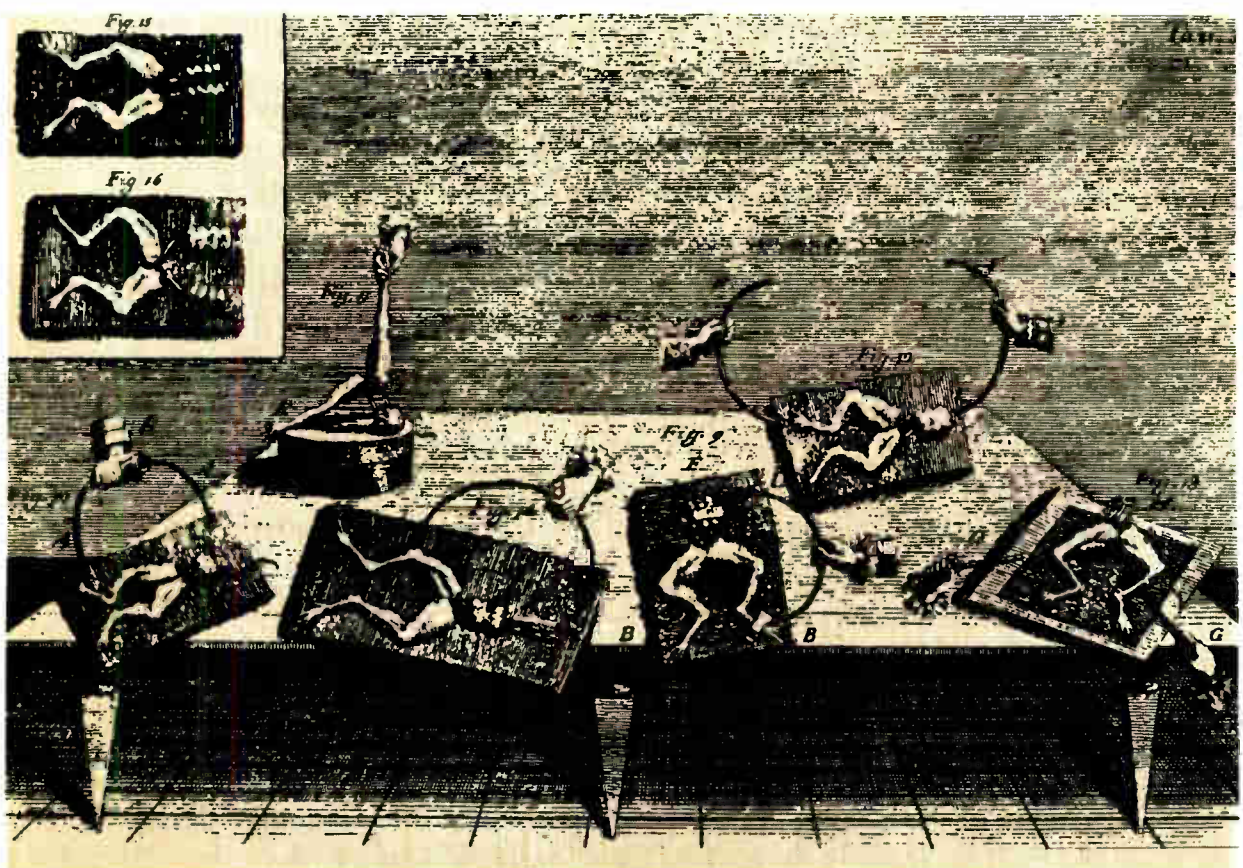


FIGURE 2. Galvani's experiment designed to determine whether atmospheric electricity (detected by the conductor, attached to the frog-leg preparation on the table) would cause contractions in the muscles. (From Ref. 2, by permission)

FIGURE 3. Galvani's second experiment. (From Ref. 2, by permission)



part to another of a nerve, or muscle, etc. as well as interiorly by their conducting fibers as exteriorly by means of metallic conductors, not a consequence of a respective excess or defect, but by action proper to these metals when they are of different kinds. It is thus, says Mr. Volta, that I have discovered a new law, which is not so much a law of animal electricity, as a law of common electricity; to which ought to be attributed most of the phenomena, which would appear, from Galvani's experiments and mine, to belong to a true spontaneous animal electricity, and which is not so; but are really the effects of a very weak artificial electricity. As to the motion of the muscles, my experiments, varied in all possible ways, show that the motion of the electric fluid, excited in the organs, does not act immediately on the muscles; that it only excites the nerves, and that these, put in motion, excite in turn the muscles."

Volta believed that the source of potential that stimulated the spinal nerves (and which could be shown to evoke sensation by placing a copper coin above and a silver coin below the tongue and bringing the edges of the coins together) was the "mere mutual contact of different kinds of metal, and even by that of other conductors, also different from each other either liquid or containing some liquid, to which they are properly indebted for their conducting power."[8] Despite the belief that only dissimilar metals were required he was unsuccessful in producing electricity with a stack of coins of two different metals. Finally succeeding with a pile that consisted of both dissimilar metals and an electrolyte, he wrote[8]:

"The apparatus to which I allude [Fig. 4] and which will, no doubt astonish you, is only the assemblage of a number of good conductors of different kinds arranged in a certain manner. Thirty, forty or more pieces of copper, or rather silver, applied each to a piece of tin, or zinc, which is much better, and as many strata of water or any other liquid which may be a better conductor, such as salt water, ley [lye], etc., or pieces of pasteboard skin etc. well soaked in the liquids; such strata are interposed between every pair or combination of two different metals in an alternate series, and always in the same order of these three kinds of conductors are all that is necessary for constituting my new instrument, which as I have said, imitates the effect of the Leyden flask, or of electric batteries by communicating the same shock as these do; but which indeed is far inferior to the activity of these batteries when highly charged, either in regard to the force and noise of the explosions, the spark, the distance at which the discharge may be affected, etc. as it equals only the effect of a battery very weakly charged, though of immense capacity: in other respects, however it surpasses the virtue and power of these batteries as it has no need, like these, of being previously charged by means of foreign electricity, and as it is capable of giving a shock every time it is properly touched, however often it may be.

"To this apparatus, much more familiar at bottom, as I shall show, and even such as I have constructed it, in its form to the natural electric organ of the torpedo or electric eel, etc. than to the Leyden flask and electric batteries, I would with to give the name the artificial electric organ and, indeed, is it not like it, composed entirely of conducting bodies? Is it not also active of itself without any previous charge, without the aid of any electricity by any of the means hitherto known? Does it not act incessantly, and without intermission? And in the last place is it not capable of giving every moment shocks of greater or less strength, according to circumstances—shocks which are renewed by each new touch, and which, when thus repeated or continued for a certain time, produce the same torpor in the limbs as is occasioned by the torpedo etc.?"

Volta soon found that evaporation of the electrolyte weakened the strength of the pile. In the same paper, he described the predecessor to the wet battery, which he characterized as a "crown of cups" (*couronne de tasses*) and which produced a stronger and longer-lasting electric force. (The Royal Society translator used the term "chain of cups.") Volta's account[8] follows. Note that he used S and Z as the symbols for silver and zinc.

"I dispose, therefore, a row of several basons [basins] or cups [Fig. 4] of any matter [material] whatever, except metal, such as wood, shell, earth, or rather glass (small tumblers or drinking glasses are the most convenient), half filled with pure water, or rather salt water or ley: they are made to communicate by forming them into a sort of chain, by means of so many metallic arcs, one arm of which, Sa, or only the extremity S, immersed in one of the tumblers, is copper or brass, or rather of copper plated with silver; and the other, Za, immersed into the next tumbler is of tin, or rather zinc. I shall here observe, that ley and other alkaline liquors are preferable when one of the metals to be immersed is tin: salt.water is preferable when it is zinc. The two metals of which each arc is composed, are soldered together in any part above that which is immersed in the liquor, and which must touch it with a surface sufficiently large: It is necessary therefore that this part should be a plate of an inch square, or very little less; the rest of the arc may be much narrower as you choose, and even a simple metallic wire. It may also consist of a third metal different from the two immersed in the tumblers, since the action on the electric fluid which results from all the contacts of several metals that immediately succeed each other ....

"A series of 30, 40 or 60 of these tumblers connected with each other in this manner, and ranged either in a straight or curved line, or bent in every manner possible, forms the whole of this new apparatus, which at bottom and in substance is the same as the columnar one [the pile] above described; as the essential part, which consists in the immediate communication of the different metals which from each couple and the mediate communication of one couple with the other, viz. by the intervention of a humid conductor exist in the one as well as the other."

Considerable time was to pass before true explanations became available for what Galvani and Volta had done. Clearly, both had demonstrated the existence of a difference of electric potential—but what produced it eluded Galvani, and even Volta. We now can identify the potential difference present in the experiments carried out by both investigators. Although Galvani thought that he had initiated muscular contractions by discharging animal electricity resident in a physiological capacitor consisting of the nerve (inner conductor) and muscle surface (outer conductor), as diagrammed in Fig. 5A, we now know that the stimulus was derived from the electromotive force that exists at an electrode–electrolyte interface; this voltage is designated the half-cell potential. It was Nernst[9] who carried out the fundamental theoretical studies that ultimately led to the measurement of

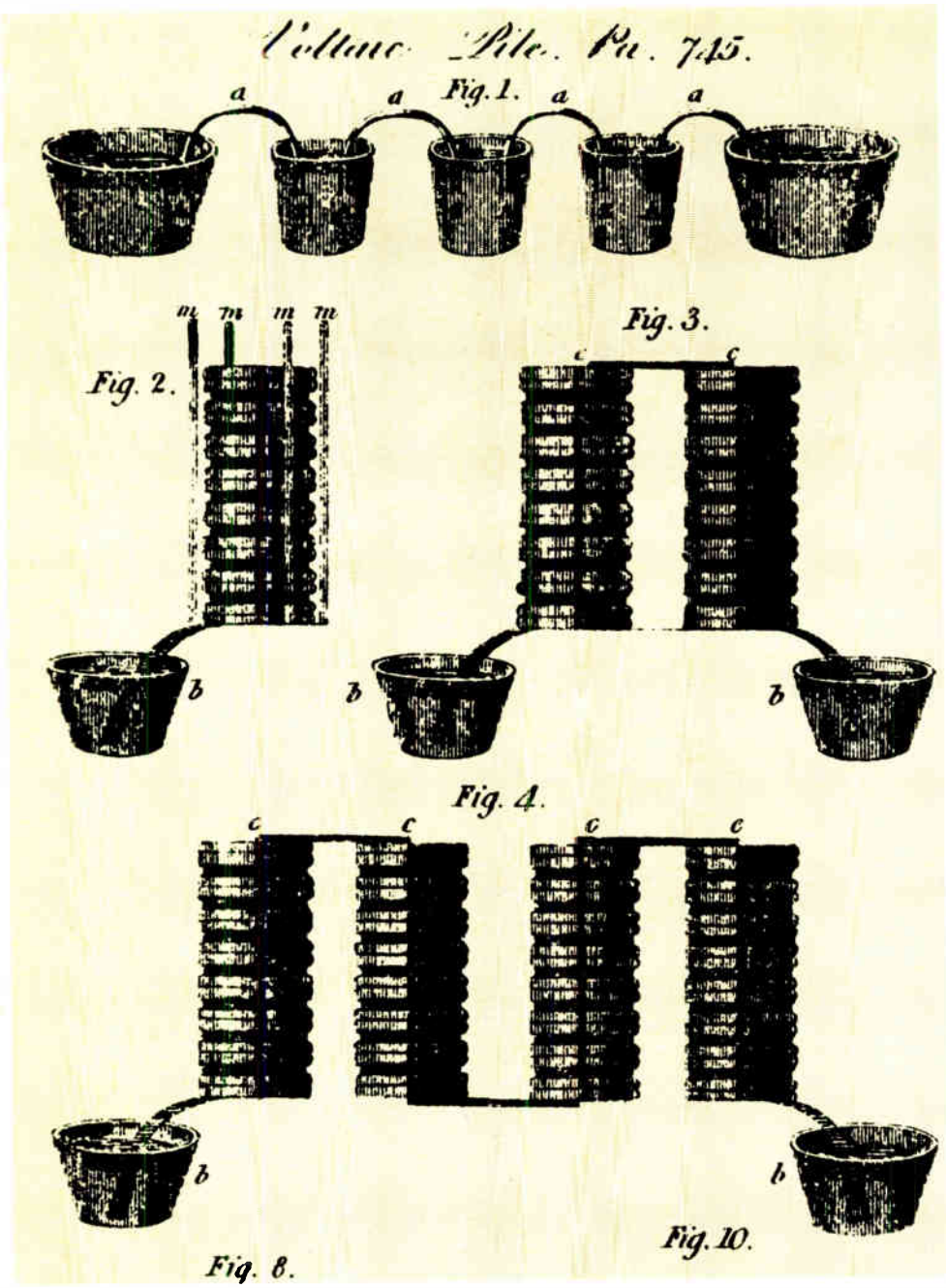


FIGURE 4. Volta's pile (Figs. 2, 3, and 4) and crown of cups (Fig. 1). (From Ref. 7, by permission)

half-cell potentials; their practical measurement was made possible by Hilderbrand's introduction[10] of the standard hydrogen electrode (SHE). Electrode potentials, measured with respect to the SHE, are found in a variety of tests on electrochemistry. It is pertinent to note that the half-cell potential of an electrode depends on the type of metal, and to a lesser degree on the concentration and temperature, of the electrolyte in which it is placed. The potential difference of a galvanic or voltaic cell, consisting of two dissimilar metals in an electrolyte of unit activity, is the difference between the half-cell potentials of the two electrodes.

With the foregoing information as background, it is possible to estimate the approximate magnitude of the potential differences encountered by Galvani and Volta. Figure 5A illustrates what Galvani thought he did (discharge a nerve–muscle Leyden jar) and Fig. 5B illustrates what he actually did—stimulate the nerve with a voltage composed of two half-cell potentials (E_A, E_B). Equating the tissue fluids of the frog to an 0.6 percent saline solution, the potential difference between iron and copper electrodes was measured and found to be 450 mV; essentially the same voltage was obtained from an iron–brass cell. Since stimulation of a nerve merely requires a small reduction (perhaps 30 mV) in the membrane potential, the voltage available from the iron–copper or iron–brass cell is ample for stimulation, even though it is not applied across cell membranes.

Volta also noted that the strength of the pile or crown of cups—as judged by an electrometer or the size of the

spark or the shock perceived on touching the ends—depended on a difference in the types of metals employed to construct the battery. (The magnetic field surrounding a current-carrying conductor, which is the basis of operation of the "galvanometer," was not discovered until 1820 by Oersted,[11] who had considerable difficulty in getting his paper published.) In a single cell of the pile or chain, he preferred silver to copper or brass for one electrode; for the other, zinc was preferable to tin. For a silver–zinc cell (assuming an electrolyte of unit activity), the potential difference is 1.56 volts; for a copper–zinc cell, it is 1.10 volts. Volta also noted that for the electrolyte, lye was better than salt water, which was better than "ordinary" water. It is interesting to note that the modern silver–zinc battery has the highest open-circuit voltage (1.86) of any of the commercially available cells, with the exception of the lead–acid cell.

Volta also found that in a series arrangement of cells, the use of a connecting wire of a different metal than those used for each cell had virtually no effect on the strength of the battery. He wrote: "It [the connecting wire] may also consist of a third metal different from the two immersed in the tumblers [containing the electrolyte] since the action on the electric fluid which results from all the contacts of several metals that immediately succeed each other, or the force with which this fluid is at last impelled, is absolutely the same, or nearly so, as that which it would have received by the immediate contact of the first metal with the last, without any intermediate metals, as I have ascertained by direct experiments, of which I shall have occasion to speak hereafter."

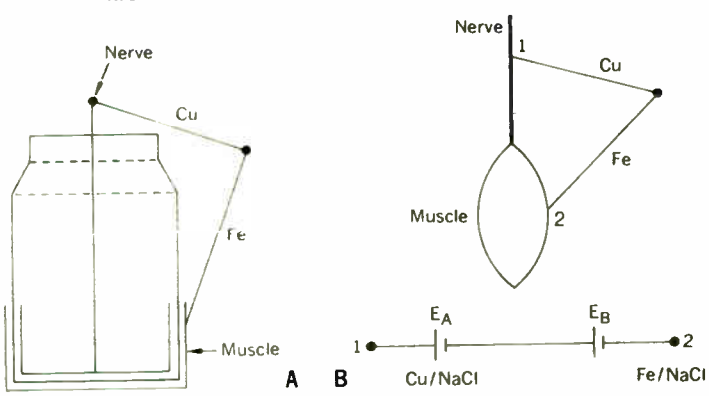
It is curious to note that Volta repeatedly ascribed the source of electromotive force in his pile and crown of cups to the contact between dissimilar metals: "The electric fluid which results from all the contacts of several metals that immediately succeed each other . . . ." "My experiments on electricity excited by the mere mutual contact of different kinds of metal . . . ." Probably this explanation was derived from experiments in which he evoked a sensation of taste, which he experienced by placing a copper and a silver coin on either side of his tongue and bringing their edges together.

Despite the inaccuracy in Volta's explanation for the operation of the electrolytic cell, it nonetheless was put to instant practical use to provide a sustained flow of current. Dibner[4] reports that Davy, for example, at the Royal Institution, constructed a battery consisting of 500 cells for his studies on the electrolytic decomposition of chemical compounds.

Volta's conclusive demonstration that Galvani had not discovered animal electricity was a blow from which the latter never recovered. Perhaps because of the death of his beloved wife, who assisted him in his experiments, and possibly also because of the troubled times in Italy resulting from Napoleon's empire-building activities, little more was heard directly from Galvani. Nevertheless, he persisted in his belief in animal electricity and conducted his third experiment, which definitely proved the existence of bioelectricity. In this experiment, he held one foot of the frog nerve–muscle preparation and swung it so that the vertebral column and the sciatic nerve touched the muscles of the other leg. When this occurred, or when the vertebral column was made to fall on the thigh, the muscles contracted vigorously. According to most historians, it was his nephew Aldini who championed Galvani's cause by describing this important experiment, in which he probably collaborated. The experiment conclusively showed that muscular contractions could be evoked without metallic conductors. According to Fulton and Cushing,[5] Aldini wrote:

"Some philosophers indeed, had conceived the idea of producing contractions in a frog without metals; and ingenious methods, proposed by my uncle Galvani, induced me to pay attention to the subject, in order that I might attain to greater simplicity. He made me sensible of the importance of the experiment and therefore I was long ago inspired with a desire of discovering that interesting process. It will be seen in the *Opuscoli* of Milan (No. 21), that I showed publicly, to the Institute of Bologna, contractions in a frog without the aid of metals so far back as the year 1794. The experiment, as described in a memoir addressed to M. Amorotti [sic] is as follows: I immersed a prepared frog in a strong solution of muriate of soda. I then took it from the solution, and, holding one extremity of it in my hand, I suffered the other to hang freely down. While in this situation, I raised up the nerves with a small glass rod, in such a manner that they did not touch the muscles. I then suddenly removed the glass rod, and every time that the spinal marrow and nerves touched the muscular parts, contractions were excited. Any idea of a stimulus arising either from the action of the salt, or from the impulse produced by the fall of the nerves, may be easily removed. Nothing will be necessary but to apply the same nerves to the muscles of another prepared frog, not in a Galvanic circle; for, in this case, neither the salt, nor the impulse even if more violent, will produce muscular motion."

FIGURE 5. Galvani's explanation for the production of muscular contraction was based on the Leyden jar (capacitor) analogy (A). He had, in fact, stimulated the nerve–muscle preparation (B) by the voltage developed by the two half-cells.



Nobili's frog current

Because this demonstration of animal electricity was so remarkable, it is important to analyze the reason for its success. We now know that even a slight injury to living cells causes the injured area to be negative with respect to uninjured regions; the potential difference between injured and intact tissue is called the injury potential and can amount to some 50 mV. Thus, application of a nerve to a muscle having intact and uninjured areas will cause the injury potential to act as a stimulus large enough to stimulate an excitable nerve. That such

an injury potential does exist, and can stimulate, was proved by Nobili.[12] He placed a skinned, decapitated frog so that its feet dipped into one glass of water or saline solution and its trunk into another, and he connected the fluid in the two glasses by a cotton or asbestos thread soaked in the solution. No metals entered the conducting arc; nevertheless, the frog muscles contracted when the connection was made. Nobili at once recognized the response as the "galvanic contraction without metals," and he accepted its origin as the result of self-stimulation of the frog by its own intrinsic electricity—the *courant de la grenouille*. He wrote:

"Having recognized the frog current so promptly, my first thought was to introduce a similar current in one of the most sensitive of my multipliers* [galvanometers], to observe the indications it would give. I had such confidence in the instrument that I employed, that I was greatly astonished to see that the frog still contracted in the new circuit while the needle of the multiplier showed not the least movement. This result made me fear that the instrument was not capable of measuring the currents produced by conductors of the second class [electrolytes], but before giving too much weight to this opinion I made a second attempt. I constructed another galvanometer, to which I gave all my attention so that it would succeed better than the preceding ones. It was in fact much more delicate, and I had indisputable signs of the frog current. The glasses being filled with common water, the deviations were of a few degrees; but with the saline solution the first movements of the needle were of 10°, 20°, and even 30°. The current of the frog goes from the muscles to the nerves, that is to say, from the feet to the head. While frogs contract under the action of their intrinsic current only for a short time, the galvanometer shows signs of it for hours."

Even though the galvanometer could indeed detect the frog current, Nobili found that a current too feeble to move the galvanometer would nevertheless evoke contractions in a frog. Studying the frog current somewhat further—and it was more or less incidental to his comparison of the physical with the physiological galvanometer—Nobili noted,[12] "The frog current has a certain direction and a certain force, and one may destroy it by directing it against another of equal intensity. If one prepares two frogs in the usual manner, and forms a galvanic circuit with them alone, by placing in reciprocal contact the nerve of one with the muscle of the other, both frogs contract; they remain motionless on the contrary, when one reverses the disposition of contacts, touching the nerve with the nerve, and the muscle with the muscle."

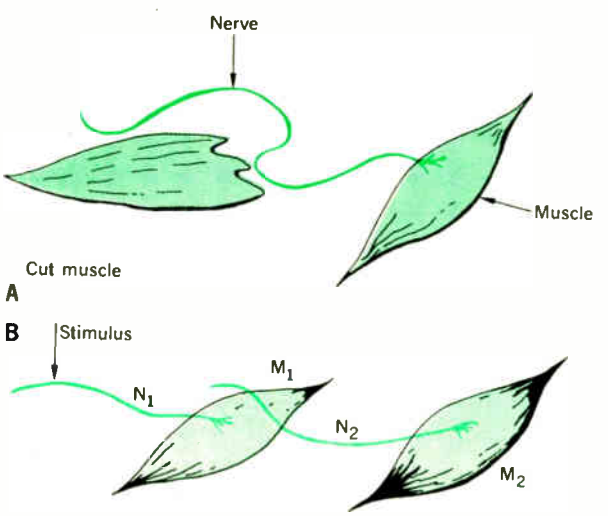
Following the reasoning suggested by these observations, he arranged frog preparations in series, placing the trunk of the second on the legs of the first, and so on, to make a "voltaic pile" of frogs, and observed an increase in the deflection of the multiplier.

Nobili thus may be credited with the first demonstration by physical instruments of the existence of the bioelectric current that today we recognize as injury current. However, he made no special point of this demonstration apart from the fact that it showed that he was able to construct an electric detector that approached, though it did not equal, the galvanoscopic frog in sensitivity. He accepted fully the evidence of the contractions of the frog as proof that the frog current did exist. Reasoning that it must arise from conductors "enough different from each other to make an appreciable current in the frog," he added, "we shall often have the occasion to talk of this current, we shall call it for this reason, *courant de la grenouille*, without bothering about the regions, whatever they are, that produce it."[12]

Matteucci's contribution

Although Galvani's and Volta's original explanations for their experiments were only partially correct, it is certainly true that Volta discovered the means whereby electric energy could be derived from chemical energy and that Galvani, by his third experiment on contractions produced without metals, proved the existence of a bioelectric potential. The subsequent history of current electricity is well known; less familiar, however, is the story of bioelectricity and its close association with and dependence on developments in the physical sciences. To summarize this story briefly, the next great contribution to the field was made by Galvani's countryman, Carlo Matteucci, who both confirmed Galvani's third experiment and made a new discovery, that of the action potential that precedes the contraction of skeletal muscle. In confirming Galvani's third experiment, which demonstrated the injury potential, Matteucci[13] noted, "I injure the muscles of any living animal whatever, and into the interior of the wound I insert the nerve of the leg, which I hold, insulated with glass tube. As I move this nervous filament in the interior of the wound, I see immediately strong contractions in the leg. To always obtain them, it is necessary that one point of the nervous filament touches the depths of the wound, and that another point

FIGURE 6. Demonstration of the injury and discovery of the action potential as shown by Matteucci's two most important experiments. A—The injury potential, which exists between an intact and an injured (cut) area, stimulates a nerve when it bridges the two regions; this is indicated by a twitch in the muscle innervated. B—Stimulation of nerve N_1 causes muscle M_1 to contract and its accompanying action potential stimulates nerve N_2 as revealed by contraction in muscle M_2.



* The term "multiplier" was used for any device that increased (e.g., multiplied) the Oersted effect; the magnetic field surrounding a current-carrying conductor.

of the same nerve touches the edge of the wound." (See Fig. 6A.)

By using a galvanometer, Matteucci found that the difference of potential between an injured and uninjured area was diminished during a tetanic contraction; study of this phenomenon was to occupy the attention of all succeeding electrophysiologists. More than this, however, Matteucci made another remarkable discovery—that accompanying a contraction of intact skeletal muscle there occurs a transient bioelectric event, now designated the action potential. He demonstrated this by showing that a contracting muscle is able to stimulate a nerve that, in turn, causes contraction of the muscle innervated by it. Matteucci[14] explained:

"I place upon an insulated surface of waxed or varnished cloth a frog, prepared in the ordinary manner; then I prepare another frog so as to have only a leg with the nervous fillet or fascicle which comes from the marrow to the muscles of the leg. It is necessary to have care, not to be led into error, to remove all the muscles of the thigh, and that the nervous filament be thoroughly denuded.

"Then I place the nervous filament upon the thighs of the first frog, in such a way that the filaments of the leg touched by the nervous filament do not come in contact with the thighs, and so that this filament is not under tension. . . . When one touches with a voltaic couple the lumbar nerves of the frog, at that instant the muscles of the thighs contract; at the same time, one sees the leg contract, whose nerve is resting upon the muscles set into contraction." Figure 6B shows the equipment.

Through the simple experiments of Galvani, Nobili, and Matteucci, the existence of a bioelectric potential was established; its transient disappearance during activity in skeletal muscle, a recording of which later became known as the action potential, was Matteucci's major contribution. Soon thereafter the presence of an action potential was discovered in cardiac muscle and nerve. Measurement of its temporal nature was accomplished by a remarkably simple instrument, the rheotome. With this device it was possible to use galvanometers that had response times far too long for the events they measured and, by utilizing what is now known as sampling theory, to obtain accurate voltage–time graphs covering only milliseconds. However, the story of these triumphs of instrumentation in the measurement of short-duration bioelectric events will be reserved for a future account.

REFERENCES

1. Kellaway, P. E. C., "The part played by electric fish in the early history of bioelectricity and electrotherapy," *Bull. Hist. Med.*, vol. 20, pp. 112–137, 1946.

2. Galvani, L., *Commentary on the Effect of Electricity on Muscular Motion* (translated by R. M. Green). New Haven, Conn.: Licht, 1953.

3. Galvani, L., *De Viribus Electricitatis* (in English). Norwalk, Conn.: Burndy Library, 1953.

4. Dibner, B., *Galvani-Volta: A Controversy That Led to the Discovery of Useful Electricity*. Norwalk, Conn.: Burndy Library, 1952.

5. Fulton, J. F., and Cushing, H., "A bibliographical study of the Galvani and Aldini writings on animal electricity," *Ann. Sci.*, vol. 1, pp. 239–268, 1936.

6. Hoff, H. E., "Galvani and the pre-Galvani electrophysiologists," *Ann. Sci.*, vol. 1, pp. 157–172, 1936.

7. Volta, A., "Account of some discoveries made by Mr. Galvani of Bologna with experiments and observations on them. In two letters from Mr. Alexander Volta, F.R.S., professor of natural philosophy in the University of Pavia, to Mr. Tiberius Cavallo, F.R.S.," *Phil. Trans. Roy. Soc. London*, vol. 83, pp. 285–291, 1793.

8. Volta, A., "On the electricity excited by the mere contact of conducting substances of different kinds. In a letter from Mr. Alexander Volta, F.R.S., professor of natural philosophy in the University of Pavia, to the Rt. Hon. Sir Joseph Banks, Bart. K.B.P.R.S." *Phil. Trans. Roy. Soc. London*, vol. 90, pp. 744–746, 1800.

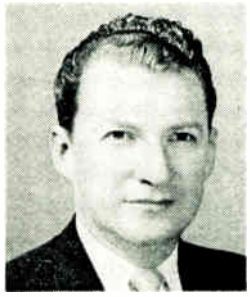
9. Nernst, W., "Die elektromotorische Wirksamkeit der Jonen," *Z. Physik. Chem.*, vol. 4, pp. 129–188, 1889.

10. Hilderbrand, J. H., "Some applications of the hydrogen electrode in analyses research and teaching," *J. Am. Chem. Soc.*, vol. 35, pp. 847–871, 1913.

11. Oersted, J. D., in *J. Chem. Phys.*, vol. 56, p. 394, 1820; also in *La Decouverte de l'Electromagnetisme Faite en 1820 par J-C. Oersted*. Copenhagen: Larsen, 1820.

12. Nobili, C. L., "Comparison entre les deux galvanomètres les plus sensibles, la grenouille et le multiplicateur a deux aiguilles, suivie de quelques résultats nouveaux," *Ann. Chim. Phys.*, vol. 38, pp. 225–245, 1828.

Reprints of this article (X71-123) are available to readers. Please use the order form on page 8, which gives information and prices.



Leslie A. Geddes (SM) is professor of physiology and chief of the Division of Biomedical Engineering at Baylor College of Medicine, Houston, Tex., and assistant professor of physiology at the University of Texas Dental College in Houston. He also is professor of physiology, College of Veterinary Medicine (Graduate School), and professor of biomedical engineering, Faculty of Engineering, at Texas A&M University and serves as a consultant to the National Institutes of Health and the National Science Foundation. Born in Scotland and educated in Canada, Dr. Geddes received the B.E. and M.E. degrees in electrical engineering from McGill University and the Ph.D. degree in physiology from the Baylor College of Medicine. He has published prolifically, is a member of many scientific and professional societies, and is currently a consulting editor to a number of periodicals, including the IEEE Transactions on Bio-Medical Engineering.

Hebbel E. Hoff is associate dean for faculty and clinical affairs, Benjamin F. Hambleton professor of physiology, and chairman of the Department of Physiology, Baylor College of Medicine, Houston, Tex. He is also visiting professor, Department of Physiology, University of Texas Dental Branch, and a consultant to the Veterans' Administration Hospital, both in Houston. Dr. Hoff holds the B.S. degree from the University of Washington, the B.A., M.A., and Ph.D. degrees from Oxford University, and the M.D. degree from Harvard University. He received the Distinguished Service Professor Award from the Baylor College of Medicine in 1968. His major research interest is the physiology of the cardiovascular-respiratory system and he has published more than 350 scientific papers.

Systems approach toward nationwide air-pollution control

III. Mathematical models

Although there is disagreement over the many approaches and designs of a pollution-control network, it seems inevitable that any lasting system must have a cost-effective basis

Robert J. Bibbero Honeywell Inc.

The future air-pollution control system that is described in Parts I and II of this article will be equipped with sensors to monitor the current status of local air pollution and meteorological factors such as wind, stability, or mixing depth. The system will be furnished not only with data on the location and strength of pollution sources and synoptic weather conditions, but with any other information needed to forecast air quality for a period of 24 hours or more. Once these statistics have been compiled, however, it is the job of the mathematical model to convert the data into pragmatic decisions for controlling the air resource.

Parts I and II of this article (Oct. and Nov. IEEE SPECTRUM) have already described the sensors and data-processing procedures necessary for the success of any nationwide air-pollution control system. Once the parameters have been established, however, to obtain optimum and purposive decisions that relate to defouling the air resource one must turn to mathematical modeling.

The symbolic relationship between the proposed mathematical air-pollution control model, its inputs, and its outputs is diagramed in Fig. 1. The model is needed in order to synthesize all of the known or assumed causal information into a meaningful pollution forecast. Knowledge of the status of receptors—those people and things affected by air pollution—permits prediction and display of the effects.

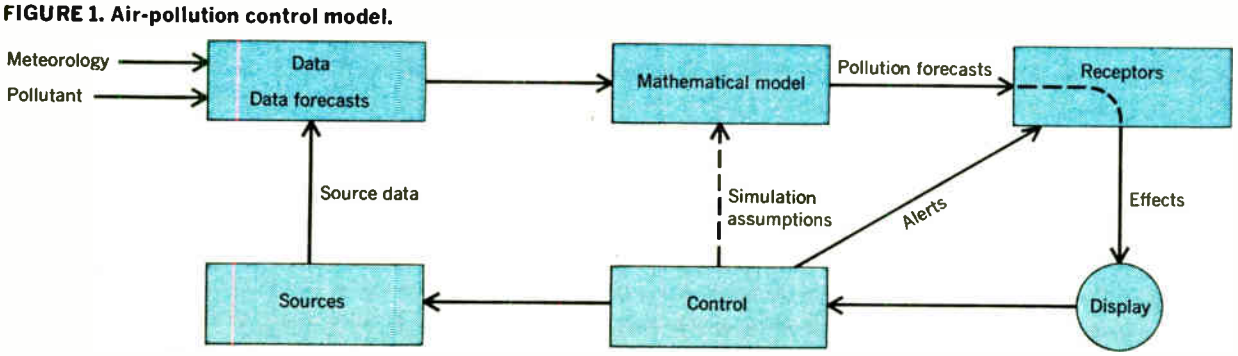
Whenever the refinement is justified, a mathematical model may also be used to facilitate the optimizing of control actions, through simulation, in order to evaluate the effects of alternate means of source control. Introduction of cost factors permits this evaluation to be on a cost-effective basis. If the model is not hopelessly complex, the decision process may be mechanized through linear programming routines. Decision simulation may proceed in real time or decisions may be predetermined off-line.

To act in these capacities, the model must be implemented in real time; that is, a solution must be obtained sufficiently in advance of the predicted effects to take corrective action. In the case of real-time simulation for control decisions, even more lead time is called for. Considering the complex causes and relationships that generate air pollution, a computerized model is strongly suggested.

Figure 2 states in more detail the input–output relationships of the model and suggests, qualitatively, its internal structure. The model may be entirely or partly empirical, based on statistical correlations (regression), or it may be analytical and derived from explicit physical relationships. A simple empirical model might be the correlation between degree-days (a measure of mean temperature for heating purposes) and SO_2 concentration averages. Generalization of this kind of relationship to places or time periods other than those for which the correlation was made would be very suspect because of the many other factors (e.g., ventilation, topography, fuel) that could influence the ambient sulfur concentration. This is true even if the correlation for a particular community, season, and averaging period is quite high. Nevertheless, such relationships at a given time may be the only kind that are available, as is now true for photochemical oxidant prediction. Such statistical relationships, if properly validated for each locality and intended use, comprise a valuable predictive tool.

An elementary analytical and physical model represent-

FIGURE 1. Air-pollution control model.



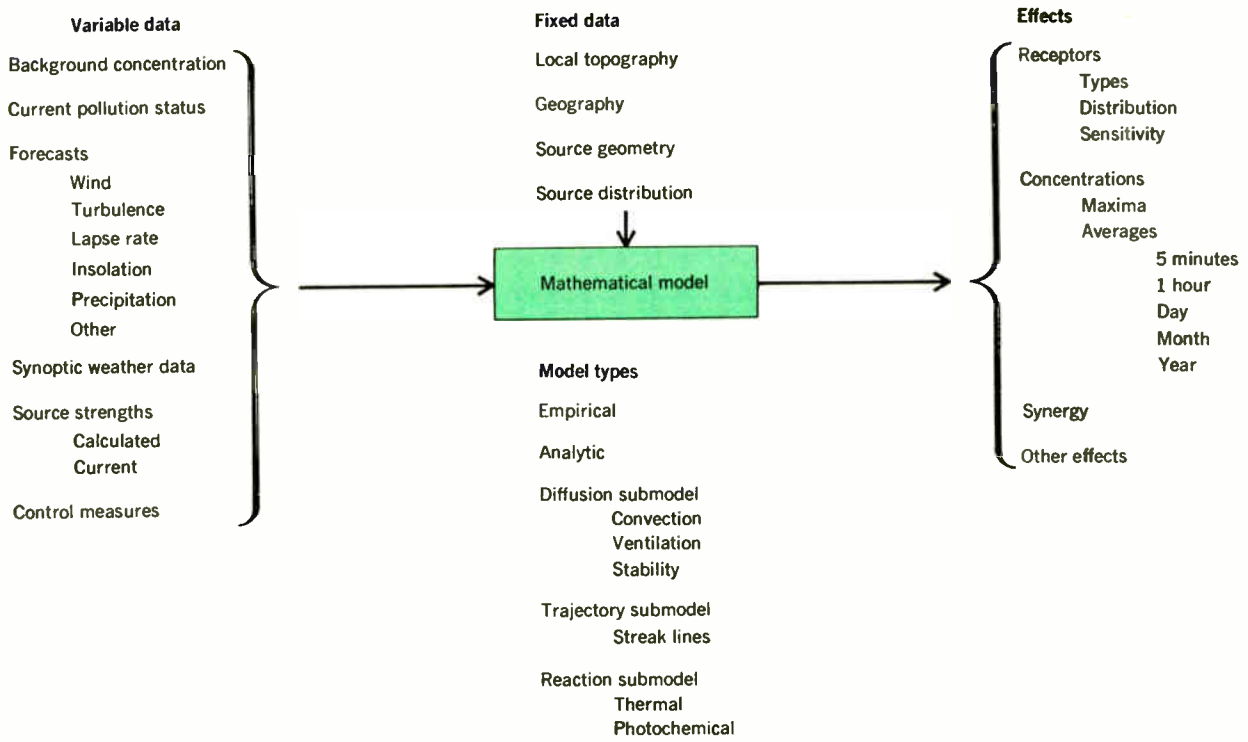


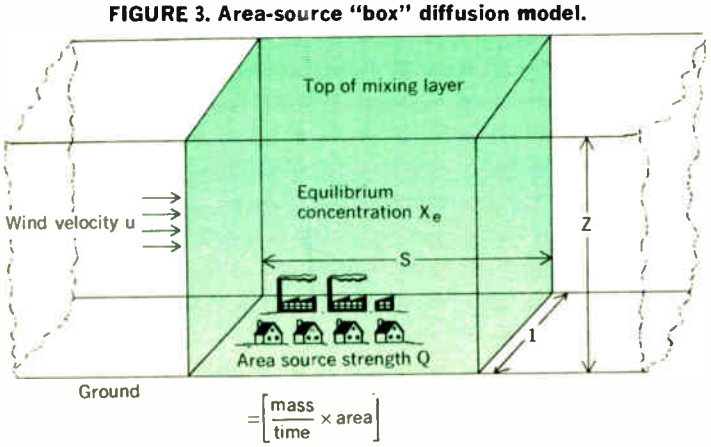
FIGURE 2. The input-output relationships of a mathematical model

ing gaseous diffusion from an area source (which may be a city with many closely spaced houses or buildings burning fuel) is shown in Fig. 3. The "box" model is considered to enclose the city, bounded by the ground (earth) as its base and the height of an assumed mixing layer Z (which may be fixed by a temperature inversion) as its top. Within the box, mixing is assumed to be thorough. The box itself is considered to be of unit width and oriented so that its length S lies in the direction of the wind, which passes freely through its ends with an average velocity u. The ventilation rate, defined as the volume of air passing through a unit width of the box, is then equal to uZ. If Q is the area source strength, or the mass emission rate per unit area, then QS is the rate for a unit width corresponding to the ventilation rate. With some mathematical manipulation, it can be shown (R. C. Wanta, Ref. 1, vol. 1, pp. 216–217; also Ref. 2) that the equilibrium concentration X_e of a gaseous pollutant in the box equals the emission-to-ventilation ratio, or

$$X_e = \frac{QS}{uZ} \quad \text{(mass per unit volume)} \qquad (1)$$

using any consistent units. Also, 90 percent of the equilibrium concentration is reached within a certain time, 2.3 S/u. For particulates, which settle with time, these expressions must be modified accordingly.

FIGURE 3. Area-source "box" diffusion model.



The box or prism model may be used for rough-order-of-magnitude assessments of a city's pollution concentration, using approximate numbers for mixing height, wind vector, and source strengths. On the other hand, it is far from answering the primary question asked of a mathematical model: "What will be the pollutant concentrations at any point in the air quality region, given all the data on sources and meteorological conditions?" In theory, a complete answer to this question can be given only by continuously tracking the pollutants emitted from individual sources and computing the concentration of each species at every point as they are transported by the wind, spread by diffusion, mixed by turbulence, and reflected or channeled by surfaces such as the ground or buildings. The basic consideration of mass continuity in fluid dynamics leads to complex vector equations that describe the time-varying changes in the concentration field. The problem is much like that of trying to describe the changing intensity of a soluble dye at every point after it is dropped into a swirling, turbulent brook, except that in our case we must also account for the chemical decay of each species with time and their reactions to each other.

The basic equations have been given in several places (Ref. 3, pp. 176–177), and after a number of simplifications, which neglect the small rates of molecular diffusion

as compared with the much greater turbulent eddy diffusion coefficients (K_x, K_y, and K_z), we are given an expression of the form

$$\frac{\partial C_i}{\partial t} + u(z) \cdot \nabla C_i = \frac{\partial}{\partial x} K_x \frac{\partial C_i}{\partial x} + \frac{\partial}{\partial y} K_y \frac{\partial C_i}{\partial y} + \frac{\partial}{\partial z} K_z \frac{\partial C_i}{\partial z} + R_i(C_1, \cdots, C_n) \quad (2)$$

In this expression, C_i is the time-averaged concentration of the species i, $u(z)$ is the average wind velocity at height Z, and R_i represents the rate of production of i by chemical reactions between species.

Since the chemical reaction term $R_i(C_1, \cdots, C_n)$ is usually nonlinear, most solutions to Eq. (2) have been limited to inert components, where the term equals zero. Even with this stringent limitation, exact solutions of (2) have been difficult to implement because of the serious problems in defining the diffusion coefficients K_x, K_y, K_z. The difficulty is that the K's depend on the size and velocity of turbulent eddies, and these in turn depend on so many factors and interactions that they become very complicated functions of their positions in the field. Many assumptions for K-values have been made by numerous sources[4] and applied to Eq. (2), but few have successfully solved the atmospheric diffusion problem other than for highly simplified situations.

Useful attempts have been made to circumvent this difficulty by utilizing the statistical properties of turbulence, rather than employing a purely analytical solution. The most popular scheme is to assume that the plume on a single species of effluent (neglecting any chemical reaction) from each source spreads out randomly as it is blown downwind, so that the pollutant concentration along any axis across the plume's cross section is distributed according to the familiar Gaussian or bell-shaped curve. This situation is illustrated in Fig. 4. The expression that describes this distribution is a form of the so-called Sutton diffusion equation (G. H. Strom, in Ref. 1, vol. 1, pp. 254–256). Given the standard deviation of concentration in the vertical and horizontal directions (the value of the vertical standard deviation σ_z is usually larger than the horizontal, σ_y), the equation can be solved for the ground-level concentration C of an inert pollutant at any distance y from a source of virtual height h. (The virtual height includes the thermal rise of the plume above the stack top[5]; Q and u have the same meaning as before.) Hence,

$$C = \frac{Q}{\pi \sigma_y \sigma_z u} \exp\left[-\frac{y^2}{2\sigma_y^2} - \frac{h^2}{2\sigma_z^2}\right] \quad (3)$$

An astute reader will observe that we have merely traded the problem of determining the diffusion coefficients K for the similar problem of evaluating the new statistical variables σ_y and σ_z. However, a number of field studies have been conducted, resulting in expressions capable of defining the standard deviations in terms of diffusion parameters, which in turn depend on the stability (lapse rate) and mixing depth as well as the gustiness of the wind. These factors will vary, of course, as a function of the terrain. Although rather extensive evaluation has been conducted at Brookhaven National Laboratories (where the ground is somewhat flat and wooded) as well as in the rolling country of Porton, England, the terminology used to describe stability and the numerical values obtained have not been in perfect agreement (see Strom, pp. 256–257). Thus there is still a problem in finding the correct numbers to put into the diffusion equations, and it would be expected that the diffusion parameters for cities would be very different from those for exurban sites.

Despite these difficulties, diffusion models have been useful. One well-established use has been to compute the maximum concentration of pollutants at any point on the ground, downwind from a stack. Equation (3) can be manipulated to give this result. If the statistics σ_z and σ_y are the same functional form of the diffusion parameters, the expression is

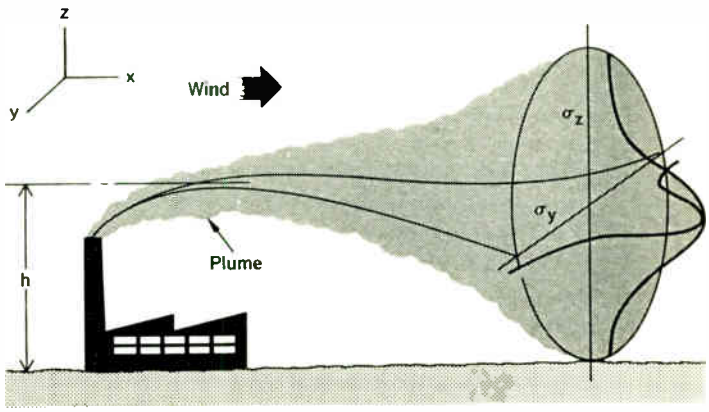
$$C_{max} = \frac{2Q}{\pi u e h^2} \frac{\sigma_z}{\sigma_y} \quad (4)$$

where e is the base of natural logarithms.

For settleable particulates, these equations must be modified. Furthermore, the role of chemical reactions, which is of overriding importance in the case of photochemical smog, has been neglected in the simple forms shown.

At this point, we should examine the results of some of the applications of these diffusion equations. Much success has been achieved using the Sutton equation and its variants to locate and design stacks for industrial power plants. Diffusion models have also been used routinely by NAPCA (now APCO) to determine the average distribution of pollutants in urban areas and to establish air quality control region (AQCR) boundaries. Moreover, at least nine separate tests of models have been conducted on a more rigorous scale in various cities. These have been reported in detail in papers by Wanta (Ref. 1, pp. 220–223) and Seinfeld (Ref. 3, pp. 178–184). In most cases, Eq. (3) or a simplified version of it comprised the basic model. On the basis of source and meteorological data, future concentrations were predicted for the pollutants SO_2, NO_x, and CO or CO_2 in some cases. The only chemical reaction considered was the decay of SO_2. Generally, the resolution of the model in space measured a kilometer or more and, in terms of time, from 1 or 2 hours to a month. Under these conditions, Pooler[6] in 1961 was able to predict half of the monthly averages of SO_2 at 123 stations in Nashville, Tenn., within a factor of 1.25. In 1964, Clark refined the time scale to 2 hours and was able to predict 24-hour averages of NO_x at 14 of 19 stations in

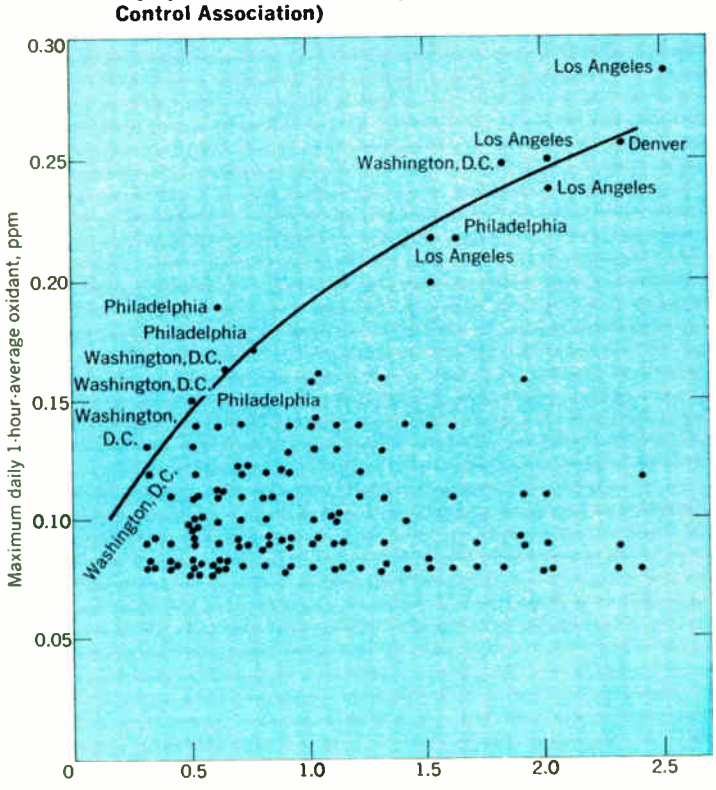
FIGURE 4. Gaussian diffusion from point source.



Cincinnati, Ohio, to within 0.02 ppm. Turner[7] reduced this error by half 58 percent of the time using Eq. (3), and in 1967 Koogler was able to report 90 percent of 8-hour SO_2 averages correctly within 0.01 ppm, employing the same model. Perhaps the most extensive tests of the Gaussian diffusion model have been conducted by Miller and Holzworth[8] in three different cities, with 2-hour-average predictions at an accuracy comparable to Koogler's.

Considering the coarseness of the input data and the resolution obtained, these results appear quite promising. Nevertheless, they are severely limited by their inability to account for the time variation of source strengths and their limitation to simple sources and inert contaminants. In the most ambitious model used to date, Lamb[9] in 1968 returned to the diffusion equation, a special form of Eq. (2), and applied it to compute CO concentration at 1200 grid points (as close to each other as 200 meters) for a single day in Los Angeles. Lamb's model utilized simple chemical reaction rates and included absorption of components by the ground. His point, line, and area sources were variable in space and time. Stability (inversion height) and the K values of Eq. (2) were considered constant. In the numerical solution of Lamb's integral equations, the sources were considered to emit a puff of pollutant at each time step: these puffs were dispersed by the x,y components of surface wind computed at each grid point, and followed until fully dispersed (Ref. 3, pp. 178–184). The effects of all these dispersed emissions were then totaled to obtain concentration as a function of time and location.

FIGURE 5. Maximum daily oxidant as a function of morning hydrocarbons. (Courtesy Journal of the Air Pollution Control Association)



The model's predictions did not correlate perfectly with measurements at various stations. Its faults have been ascribed to lack of a vertical wind component, giving concentrations too high at the convergence of trajectories and results too low during the afternoon, suggesting an influx of sources from outside of Los Angeles. In order to improve this model substantially it will also be necessary to include nonlinear chemical-reaction terms (Ref. 3, pp. 178–184). It has been reported that such a model is under development at Systems Development Corporation under the sponsorship of APCO.[10]

In practice, chemical reactions in the atmosphere have been taken into account by means of the purely statistical or correlation models mentioned previously. In general, the polluted air mass over a city is unstable chemically as well as physically. Both thermal and photochemical reactions between atmospheric contaminants occur. Thermal reactions include the formation of acid mists from SO_2 (catalyzed by oxide particulates) and salts from acids and metals, etc., some of which may be promoted by the surface effects of particulates. The most troublesome reactions are photochemical. These result in the production of oxidants through ultraviolet irradiation of certain reactive hydrocarbons mixed with nitrogen oxides.

Laboratory studies of polluted atmospheres have given sufficient information about these reaction rates to construct a simulation model. But, as noted previously, it has not been possible to integrate this into an analytical diffusion model. However, it is known that an empirical relationship exists between the concentration of hydrocarbons (nonmethane) in early morning hours and the maximum hourly average oxidant concentration that may occur later that day. Since it is also known that many other factors, such as nitrogen oxide concentration, sunlight, and meteorological ventilation, can and do intervene, the relationship between these two factors is a statistical one,[11] as seen in Fig. 5. Though limited and unsatisfactory, such data may be put to use in a negative way. That is, given the morning hydrocarbon reading it can be predicted that the maximum level of oxidant will *not* exceed some upper limit, defined by the limits plotted in Fig. 5.

Furthermore, the data can be manipulated to express an air quality standard. If the maximum allowable daily one-hour-average oxidant value is 0.1 ppm, for example, then the 6–9 A.M. average nonmethane hydrocarbon concentration must not be allowed to rise above 0.3 ppm.

Requirements for data processing

Ultimately, we hope to solve mathematical models similar to that described by Eq. (2) in sufficient detail to permit forecasts of air pollution within urban regions. These forecasts must be computed rapidly if the information is to be usefully applied to preventive control and management of pollution sources. Even if we know the K values, the chemical reactions and their rates, and the characteristics of wind and turbulence around urban areas and buildings, the simultaneous numerical solution of the many partial differential equations is formidable. Though we do not know the exact nature of the ultimate model or what simplifications may be introduced, it is legitimate to ask what order of processing capability may be demanded.

If the Gaussian diffusion model represented by Eq. (3) were to be used, this question could be answered with

some confidence. Turner[7] computed 24-hour forecasts for Nashville with this model, employing 2 minutes of time on an IBM 7090 machine. It is possible to scale his methods to a hypothetical area the size of New York City by assuming 200 grid points rather than the 99 used for Nashville, four topographical levels instead of one, and a doubled time resolution. This yields a progressing time multiplier of 16. The IBM 7090 cycle time is 1400 ns; a typical modern minicomputer such as the Honeywell DDP-516 cycles at 960 ns, a 1975 mini may cycle at 750 ns, whereas a large 1975 processor may cycle as fast as 40 ns. Of course, it is not possible to compare the processing time of "benchmark" problems by cycle time alone, since the total machine architecture and software organization equally influence this parameter. But if we assume that all these factors vary proportionally, the processing time for the hypothetical New York City prediction problem would be

Present minicomputer (e.g., Honeywell DDP-516)	22 minutes
1975 minicomputer	17 minutes
1975 large processor	< 1 minute

Assuming the ability to compute predicted pollution concentrations, the next step is to compute optimum control measures by evaluating alternate control strategies. An approach to solving this problem has been made by applying a linear programming routine.[12] This was applied to a model of the St. Louis air shed and allowed 200 control measures (such as fuel switching, leaf collection, automobile and process controls, etc.) to be applied against the constraints of availability, maximum effectiveness of each measure, and total requirements for pollution reduction. The model accounted for the five major pollutants and a large number of sources (unspecified in Ref. 12). The output of the computation was a set of control methods that eliminated the desired weight of pollutants from the air shed, at the least total cost.

Although this model is useful as a tool for air-pollution control and is simple enough to be applicable generally to other air sheds, it does not contain enough detail to determine localized or neighborhood effects or to solve short-term problems. Correction of all these limitations would produce a much larger and more sophisticated model. The present one was programmed for an IBM 360 computer. Although the running time was not given, it can be guessed that an optimization problem might run to the equivalent of 100 iterations of the Gaussian diffusion model. On this assumption, the resultant running time for the assumed 1975 large processor discussed will be of the order of 1½ hours. It is probable, therefore, that optimum solutions for various hypothetical pollution situations will be predetermined and stored for future use, or that suboptimal shortcut procedures will be developed.

The computation picture based on solutions to the inert Gaussian diffusion equations may be far too optimistic if the analytic diffusion model is implemented. An estimate[13] for one such complicated model runs to one hour computing time per day of forecast on a large parallel processor such as Illiac IV. Although machines of this power might not be in abundance in 1975, they will certainly be available and it has been suggested that one may be time-shared; thus one machine can service 20 or more cities.[14]

The need for real-time data has been discussed earlier, but before concluding this section, it should be conceded that opinion on this subject, relative to existing monitoring networks, is by no means unanimous. Mitre Corporation recently completed a study of this topic based on interviews with federal and state air-pollution officials.[15] It was found that only 28 percent of them believed it necessary to have a national monitoring network capable of reporting in less than 6 hours, whereas 62 percent thought it necessary at the regional and local level. Those in favor of real-time data cited episode and source control as their primary needs, mentioning medical (alerts to hospitals and patients) and industrial accident control as well. Those opposed to rapid air-pollution data dissemination apparently believed so because of the lack of effective control action that can be taken in a short time, especially at the federal level. Since this attitude realistically reflects the situation in many cases, it can be expected to change as the legal and technical steps for source control are taken.

Even in today's networks, however, the study conceded the need for telemetering of warning signals (high concentrations) from remote, unattended stations and the need for large processors to manage data at central stations. It also foresaw the need for data exchange between local networks to combat interregional or national pollution transport, and for this reason a uniform data format compatible with the federal climatological network is recommended. (SAROAD is such a code.)

Source control and abatement

At the operative end of the nationwide air-pollution system is the source-control subsystem, and its ultimate technical objective is to correct the predicted adverse trends in atmospheric pollution. The word "control" is used here in two senses, meaning both the legal emission standards and ordinances reflecting the social power over pollution-emitting sources and the physical devices that actully reduce emission. We shall discuss mainly the concepts and means of implementing the legal standards, including measuring devices at the sources to see that these controls are enforced.

Although the physical controls are of immense technical and monetary interest to industry, representing perhaps 90 percent of the funds that must be expended to clean the air, they represent little that is not well known. Much of the expenditures for air-pollution control will go for such mundane devices as bag filters, cyclones, gas scrubbers, and absorbers. Electronic precipitators for removing particulates, especially, are immensely expensive, as mentioned earlier, and of somewhat greater technical interest, but have been used since the beginning of this century. Tall stacks, of course, do not reduce total emissions, but according to Eq. (4) they will reduce the concentration of ground pollution by the square of their height; however, they are no less expensive, costing from $3000 to $7500 per meter. The greatest area for innovation may be actual changes in the pollution-causing processes themselves. If materials can be used more efficiently or recycled, the end result may be a saving rather than an expense to the manufacturer.[16,17]

Legal control may be exerted prior to or following the establishment of a man-made source of air pollution. Physical controls and tests to prove the ability to meet emission standards may be required as a prerequisite for construction or operation. Once a source is permitted to

Spacing of Lines on Ringelmann Chart

Ringelmann chart no.	Width of black lines (mm)	Width of white spaces (mm)	Percent black
0	All white		0
1	1	9	20
2	2.3	7.7	40
3	3.7	6.3	60
4	5.5	4.5	80
5	All black		100

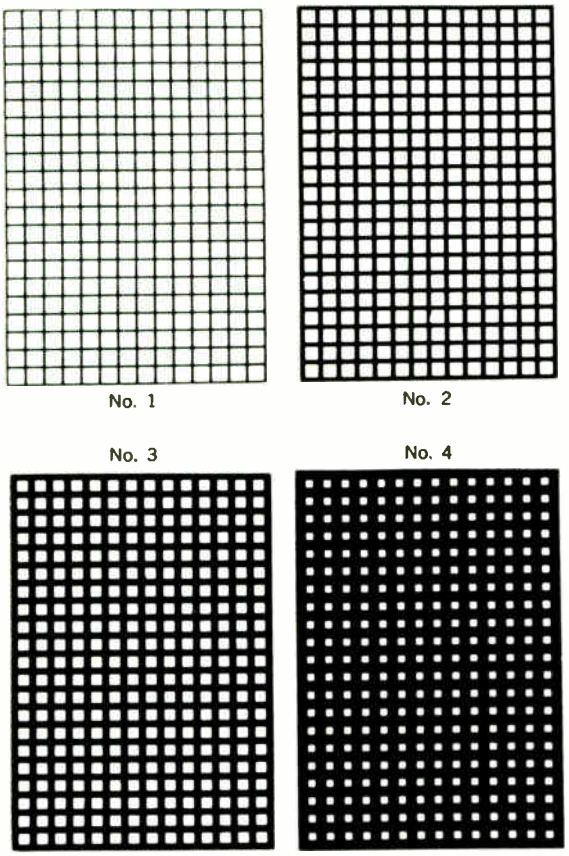
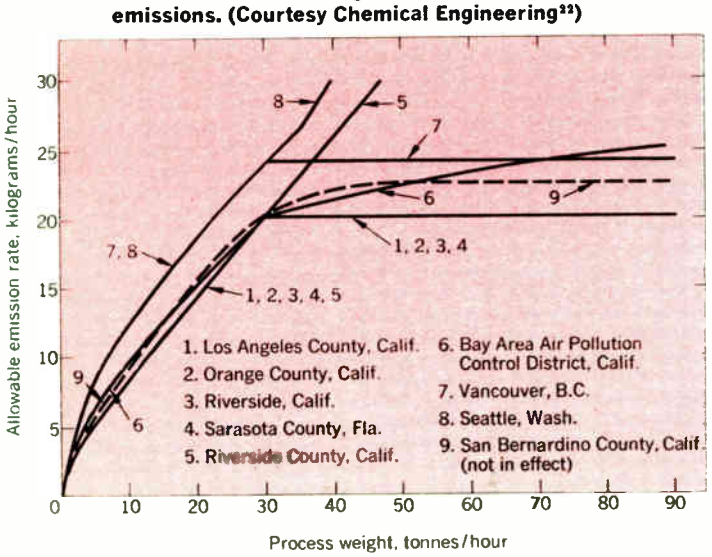


FIGURE 6. The Ringelmann chart. (Courtesy Academic Press)

FIGURE 7. Process-weight emission rule for particulate emissions. (Courtesy Chemical Engineering[22])



come into existence, its emissions must be brought into line with the regional air quality standards by continual monitoring and inspection. If it is found wanting in this respect, additional controls or a change in process may be required.

Prior controls

Ideally, air-resource management starts with community planning. All aspects of land-use planning—such as transportation; zoning for industry, commerce, or residential use; waste disposal; and park and open-space reservation—should be considered in planning to meet quality standards for the air, and conversely. A survey of the local background of contaminants should be made prior to land development. Such factors as future growth of population, increased automobile use, location of new electric power plants, and introduction of new mass transportation methods have obvious impact on future air quality. The effect of these changes as well as the expected values of pollution with current land usage can be employed by the use of mathematical diffusion models, as previously described.

Given a single source with a constant rate of emission and stack height, there is a combination of wind direction and velocity, atmospheric stability, and distance for which the ground concentration of pollutant will be greatest [Eq. (4)]. This concentration can be numerically set equal to the desired air quality in order to obtain an emission standard for this stack. In the case of several large and many small sources, the problem of allocating the total emission may become very complex, but this is precisely the basic problem of fairly dividing up use of the public's air resource. (Croke and Roberts,[18] for example, have suggested emission standards based on land area owned, rather than by individual stack or industrial plant.) A computer program that has for inputs the yearly inventory of source strengths and a long-range analysis of meteorological variables, with the capability of computing the time and place of maximum pollution, has been suggested by Stern (Ref. 1, vol. 3, p. 620). The program would then print out the contribution of each source to this "worst pollution day" and permit allocation as before.

The means by which the sources in a region can control their emissions to predetermined maximums may be optimized from a cost standpoint by the use of linear programming models and similar economic models discussed earlier. Studies of this nature will lead to a rational basis for land-use and zoning decisions, and for rules governing construction permits and licenses. Where local interests conflict with quality standards based on public health and welfare, an educational program or intervention by a higher political entity, through the imposition of state or federal emission standards, may be required. Likewise, these may be necessary in the event of pollution transport between local regions.

Emission standards and zoning regulations, established by local ordinance or statute, imply the existence of a local control activity to implement them. Engineering analysis of plans, with special attention to emission points, stack heights, process-flow sheets and quantities, and proposed control methods and monitoring instruments or access, form the rationale for approval of construction permits. At this time, the data base for a source-emission inventory is established. A data bank, which may or may

not be mechanized, depending on the needs of the system, is part of the activity.

Posterior control of stationary sources

Little or no control can be exerted over stationary air-pollution sources, once they are established, without the specific legal and financial authority necessary to create and support an enforcement agency to act at the local level. The President's Council on Environmental Quality, in its annual report to the U.S. Congress,[19] indicates that most of the state programs, as well as those at the local and regional level, are inadequate in this respect. There are 144 local agencies receiving federal grants, but only 44 percent are adequately funded for a minimal program, as are only six out of the 55 funded state and territorial programs.

Prior to August 1971, there was in effect only one federal emission standard—that established for new automobiles—although performance standards for certain categories of new stationary plants have now been published.[20] In 1970, all states had air-pollution legislation, but only 42 actually had any kind of regulation to control emission. For example, 33 states had open-burning regulations (compared with 19 in 1968) but only six states regulated vehicle emissions.[21]

It is clear that the legal basis for regulation and control of sources, though accelerating, is still very inadequate and spotty. Nevertheless, patterns of source-control regulations have been established in those areas where they have been pioneered, and it is believed that new regulations for stationary sources will continue to follow these trends.

Particulates. Regulations for the control of particulates have been based on four different concepts: opacity, concentration, process weight, and emission potential.

The classic standard that is applied to the control of black smoke, and the only regulation in many communities, is that of the Ringelmann chart, first introduced in 1897 (see Fig. 6). It is a subjective measure, performed by visually comparing black smoke plumes with four cross-hatched charts exhibiting various proportions of black line width to white spacing. The charts are stationed at such distance as to merge the black lines into a uniform gray. The basic concept was expanded, first in Los Angeles, to include the "equivalent opacity" of white and colored plumes, a comparison that is even more subjective. Nevertheless, it has been possible to train inspectors to agree within a half Ringelmann "number," and the method is still firmly entrenched in enforcement practice despite its obvious shortcomings.

Other particulate regulations specify the concentration (mass per unit volume) of effluent gas. These are based on a "model smoke ordinance" developed by the ASME in 1948. Typically, the limits range from 0.2 to 0.3 grain per standard ft^3 (7–10.5 per m^3) (at 60°F and 1 atmosphere), depending on the definition of particulate, the sampling method, and the gas composition.

Another rule governs the emission of dust as a function of the weight of material processed in order to circumvent attempts to avoid the concentration rule by diluting the gas stream. The process-weight concept is demonstrated in Fig. 7. It can be seen that permissible emissions under this rule can be increased by using two or more small units, rather than one large device.

The concept of control of both gases and particulates by "emission potential," depicted in Fig. 8, has been adopted by New York State. The abscissa of each curve represents the potential rate at which contaminants would be emitted if there were no gas-cleaning devices. The rules increase in stringency depending on the toxicity class of the material; thus, about 15 percent of iron oxide potential can be emitted but only 1 percent of beryllium.[22]

It is clear that this type of standard, recognizing the effect on the receptor, is most rational. Some regulations go even further by distinguishing between fine and coarse particles, which have different physiological and physical effects. It is possible that the development of a convenient, accurate, objective instrument, which would correlate with the more important receptor effects, would create a new standard means of measuring particulates and replace the visual tests.

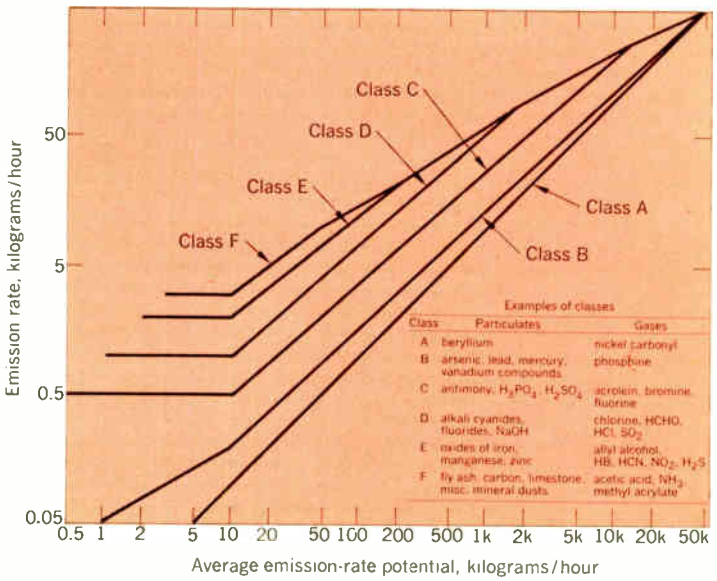
Gases. Most standards for the emissions of gases from stationary sources have been directed against SO_2, although other gases and vapors, including fluorides, hydrocarbons, solvents, and other sulfur compounds, are also controlled. In California, the rules give an operator the option to monitor ambient ground-level concentrations, rather than adhere to a fixed stack concentration limit. The ambient limits (Fig. 9) must be monitored by at least three continuous SO_2 analyzers and one recording wind station.[22]

Some emission standards are combined with a design standard, such as stack height or adjusted height (corrected for temperature of flue gas). Stack height criterions result from the diffusion models discussed earlier. Other regulations control emissions indirectly by specifying fuel standards; for example, the volatile content of coal, the olefin content of gasoline, and the sulfur content of heating fuels.

Source testing and monitoring

If, in the long term, the application of air-pollution controls to old and new sources is the response of the national system to intolerable ambient levels, then source testing and monitoring represent the feedback that closes the control loop. Source emissions are tested by the plant

FIGURE 8. "Emission potential" standard. (Courtesy Chemical Engineering[22])



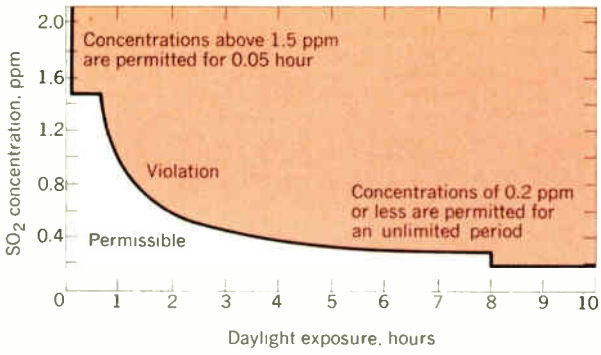


FIGURE 9. Ambient ground-level limits on gaseous emissions. (Courtesy Chemical Engineering[22])

operator or by a control agency for many reasons: survey, licensing or inspection, checking the efficiency of collectors, monitoring process malfunctions or accidents, compliance with legal standards, or for record. Such monitoring is now usually accomplished in the stack, by roof monitor, or at other effluent locations. Control agencies may, in addition, monitor or test sources with in-stack or remote sensors to answer complaints or to detect violations. In connection with alleged offenses, the method must meet local legal criterions.

Source testing must take into account all the technical problems of obtaining a valid and representative sample. Pertinent factors include cyclic or random fluctuations of the effluent, both in quantity and composition, and physical or chemical instability of the sample. Each problem must be solved in the context of the particular process, plant, and installation. The temperature and the dew point of flue gases are significant. In general, particulates pose greater sampling problems than gases because of agglomeration and variations in particle size.

Flow rates, as well as composition, directly determine a plant's contaminant emissions. Flow measurement by means of Pitot tubes in ducts and stacks must cope with plugging by particulates. Locations of sample or duct traverse points must follow good practice for these instruments. Because of the dirty, hot, or corrosive nature of effluents, special sampling equipment must often be designed. Probes and nozzles may have to be heated to prevent condensation of the sample before reaching a collector. In sampling particulates 3 μm or greater in size care must be taken to sample "isokinetically"; that is, to match the velocity of the sampling nozzle with that of the gas stream. Otherwise, an imbalance of heavy or light particles enters the nozzle.

Accurate sampling of a large duct, ensuring the proper number of sampling points, changing nozzle sizes for isokinetic conditions, flow metering, and so forth, can take 10 or more man-hours per trial. The samples so collected must be tested in the laboratory, which involves additional time and expense. Therefore, approximate or automated methods are indicated.

Continuous analysis is required to meet the more stringent regulations. The time constants of analysis equipment such as the SO_2 conductimetric monitor are short enough to gather real-time data for the guidance of operators. Photoelectric monitors installed in the stack give the same kind of information for particulates. On the other hand, some monitoring equipment, such as tape samplers, requires a definite sample period and average over this time lag. Monitoring devices relying directly on pollution effects may have very long time constants; examples are lead sulfate candles, lead acetate tiles, rubber strips, paint, and metal. These may take months or a year to measure contamination.

In general, the stack-mounted analysis equipment in current use is similar to the air quality monitoring equipment discussed earlier, except that the former measures higher concentration ranges. Wet-chemical analyzers, including both photometric and conductimetric systems, are used for SO_2 and H_2S. Flame photometric techniques are also used. Carbon monoxide is monitored by nondispersive infrared instruments; particulates, by photoelectric and infrared opacity meters and tape samplers. Recorders, and sometimes level alarms, are usually incorporated in these instruments.

A novel method of stack sampling for SO_2 is the correlation spectrograph, mentioned in Part II as a long-path sensor. The instrument measures the intensity of the ultraviolet spectrum of stack gases sampled by a slotted tube in the stack, comparing it with a photographic mask of the SO_2 spectrum. The photocell monitor is calibrated to read in parts per million, and is protected from the flue gases by means of an air curtain.

Although modern photoelectric and spectroscopic devices may solve many of the problems of source monitoring, equipment capable of reading stack-emission compositions at a distance is even more attractive. In this way, not only are installation and sampling problems obviated, but the possibility of time sharing as well as portable application for law-enforcing agencies is introduced. Remote-reading stack monitors are by no means a new idea. The Ringelmann smoke chart is a remote comparison device, and several other portable visual smoke guides and viewing devices are available.

To replace visual particulate measures, a pulsed-laser instrument (lidar) has been developed by the Stanford Research Institute and is undergoing evaluation. The Mark V acts similarly to radar, utilizing a pulsed, Q-switched ruby laser with a beam width of 0.35 mrad.[23] Reduction in light backscatter due to the smoke plume is the measurement criterion, since the echo is a function of particle size and color as well as concentration. More advanced lidars have been built and tested.

Active, single-ended, spectrographic devices, not dependent on sunlight as a source, are being studied. The infrared backscatter instrument has also been mentioned; another concept monitors emission spectrums from hot stack gases by means of a spectrophotometer. Raman spectroscopy is being investigated for stack monitoring, as referenced earlier. The difficulty with these remote systems at present, aside from the need to prove their feasibility, consistency, and acceptability as legal standards, is their cost and complexity. Less spectacular remote means, including time-lapse photography, are coming into use pending further developments.

Another requirement would involve a simple means to track down violators of pollution regulations and to locate offending sources. Tracking small windborne balloons at night, when an odor plume travels under an inversion, is one technique. The tracker must first locate the point of strongest detectable odor, release and track the balloon, and record the wind direction. Two or three

such measurements can pinpoint the odor source by triangulation. More direct means, such as the remote sensors discussed, are clearly desirable.

Short-term and episode control

Federal law requires AQCR plans to provide adequate authority to deal promptly with emergency air-pollution conditions, and some areas have so defined alert concentrations and procedures. The governors of most states have the power to take action in the event of dangerous air-pollution episodes. Alert plans call for action to be taken upon observing high concentrations of certain pollutants or combinations for as little as an hour or even less (see Table I). The designation of greatly increased numbers of air quality regions and implementation of their approved plans implies that a large number of areas will have the capability to respond within hours or less to transient but dangerous air-pollution levels.

In the present stage of development of the national air-pollution control system, however, the pace is generally much more leisurely. All regions have not implemented alert plans. The need for rapid transmission of air-monitoring data to a central data center is minimized or even doubted by some officials, as has been discussed. There have been, fortunately, few episodes in this hemisphere where widespread fatalities and illness can clearly be laid to air pollution, so that no "clear and present danger" is generally feared.

Nevertheless, the factors referenced in Part I of this article, including increased population, greater industrial activity, and more automobiles, are pointing to higher levels of contamination concentration over a *longer-term period*, while the criterions, especially those based on health effects, move toward recognition of lower levels of tolerance.[24] (It is allowed that a temporary respite, perhaps through 1980, may be obtained by stringent enforcement of federal automobile and stationary-source controls. Perhaps by that time we will achieve more basic means of restraint.) Consequently, potentially dangerous episodes should become increasingly prevalent through both increased base concentration levels and statistical fluctuations, as well as by more rigorous definitions. It follows that the mature national air-pollution control system shall have a fully developed capability to predict, detect, and react to dangerous air-pollution concentrations, acting within a time scale of hours or fractions of an hour.

In addition to this capability, which will be designed to combat the rare, natural coincidence of source-strength fluctuation and unfavorable weather, the high-speed response of a monitoring system will be effective against industrial accidents and violations of control regulations.

The quick-reaction system envisioned will depend very much on high-speed data-processing equipment and refined mathematical models that will permit prediction and prevention of dangerous concentrations, not merely their detection after the fact. Furthermore, the system will depend on a more widespread network of monitoring instrumentation than is currently deployed (Ref. 24, p. 38), and may utilize continuous central monitoring sources through permanently installed telemetering and measuring instruments as remote-acting stack monitors.

Equally important, a quick-reaction system must have a fully developed command and control structure in as real a sense as that of a military air-defense system. Information must not only be collected, but displayed without delay to personnel capable of making rapid decisions (with computer aid) on the course of action that will avert the predicted emergency.

It is understood, then, that although some portions of the envisioned system may now exist in the more highly developed air quality control systems, both here and abroad, all the necessary components do not and cannot exist until further research and development are carried out. An outline of the functions of such an ideal system is presented in Table II. These can be compared with the characteristics of some existing systems described in the next section.

Current system status

An overview of some continuous aerometric networks currently being used and developed by state and local

I. Alert stages for toxic air pollutants —Los Angeles County (parts per million)

Gas	1st Alert	2nd Alert	3rd Alert
CO	100 for 1 hour	100 for 2 hours	—
	200 for ½ hour	200 for 1 hour	200 for 2 hours
	300 for 10 min	300 for 20 min	300 for 1 hour
NO_x	3	5	10
SO_x	3	5	10
Ozone	0.5	1.0	1.5

II. Outline of quick-reaction system functions

1. Alert decision
   A. Resulting from predictions of probable dangerous pollutant levels, a decision is made to exert control over emission sources. [Note: Existing systems (e.g., Los Angeles) alert on the basis of current rather than predicted levels, plus meteorological forecasting of pollution potential in the New York–New Jersey area; see Table I.]
   B. Requirements for decision function are
      (i) Prediction model.
      (ii) Law and doctrine.
      (iii) Display of current and predicted levels, including meteorology and geographic display.
      (iv) Mobilized command and control organization.
2. Tradeoff decision
   A. Objective is to determine the least costly effective control action.
   B. Control actions available by law to the control team are
      (i) Point source abatement or shutdown.
      (ii) Stationary area source abatement (heating, process combustion, open burning, etc.).
      (iii) Mobile source abatement (e.g., auto traffic diversion or reduction).
   C. Determine the minimum cost objective according to the cost-effectiveness model or doctrine.
3. Command and control
   A. Objective: to exert control over emissions.
   B. Requirements:
      (i) Data central and status display.
      (ii) Personnel who are authorized to take competent action.
      (iii) A communications net to controlled sources. Also a public information net to control area emissions, give out information, etc.
4. Feedback
   A. Quality monitoring and trend prediction.
   B. Violation detection.
      (i) Patrol with mobile remote sensors.
   C. Source instrumentation and telemetering.

III. Continuous aerometric methods

	System	New York State	Pennsylvania [Planned]	New Jersey	Delaware
Mission	Coverage	[725 linear km]	[10 air basins]	[State]	[State]
	Population ($\times 10^6$)	—	—	—	—
	Major system objectives	monitor/alert/criterions	monitor/alert/criterions/control/ model development	monitor/alert/ criterions	—
Sample system design	Variables: Chemical	8	7	10	7
	Meteorol.	8	5	7	2
	Poll interval (minutes)	15	1	—	3
	Basic averaging time	15 min	1 hour	15 min	15 min
	Number of stations	11 [50]	1 [25]	18 + 3 mobile	4
	Sample density (stations/2.6 km²)	—	—	—	—
	Site design (primary criterions)	geography/pop. levels	diffusion model/10 air basins	—	—
Data handling	Data transmission	digital/dial-up line	digital/leased voice line	analog	—
	Data processor	B-3500	Spectra 70 + control computer	Spectra 70-45 + PDP-8	—
	Display	CRT + TTY	graphics map + printout	printout	printout
	Mathematical models: Episode prediction	—	[diffusion models]	—	—
	Control optimizing	—	—	—	—
Control decision	Alert criterions	SO_2/CO/particulate levels	—	—	—
	Control procedure	—	—	—	—
Auxiliary	Emission inventory	—	[computerized]	—	—
	Data-exchange format	—	—	—	—
	References	25	26	27	28

Note: Square brackets signify plans or goals.

air-pollution control activites will give some idea of how far we are along the road toward a nationwide monitoring and control function. This survey will disregard the National Air Sampling Network (NASN) and the Continuous Air Monitoring Program (CAMP) of the federal government, despite the fact that these *are* nationwide, because these projects are intermittent or limited in scope. (CAMP has only six stations in as many cities and is intended to provide historical and research data, rather than to serve the "real-time" needs of public information and source control.)

Several typical state and local continuous monitoring networks are described in Table III, with the Rijnmond (Rotterdam) network, developed for the Netherlands government, included for comparison.

Network characteristics can be conveniently broken down into four groups for purposes of this tabulation. These are the network mission, the sampling design, the data-handling methods, and the control action decision. The actual performance of control activities, such as fuel switching, are not part of the aerometric network, but the compliance of sources on a legal or voluntary basis is, of course, essential to the entire undertaking. The network may also take on supplementary functions, such as storage of source-emission inventories and of data in standardized format (SAROAD) for nationwide exchange.

The objectives of the networks are reasonably consistent. All have, as one goal, monitoring for unhealthy concentrations, which is not surprising in view of the requirement to establish an alert procedure in order to qualify for federal funds. (Most network funds, for example two thirds of the city of Philadelphia's total, are derived from the Federal Clean Air Act.) There is a strong tendency to spread out the objectives to longer-

Allegheny County (Pittsburgh)	Chicago	Los Angeles County	New York City	Philadelphia	Rijnmond (Rotterdam)
1100 km² 1.6	650 km² 6.2	1035 km² 6.9	520 km² 11.4	325 km² —	325 km² —
monitor/control/ historic data	monitor/alert/ control	monitor/alert/ control	monitor/alert/ criterions/hist.	monitor/alert	alert/control
7 4	1(SO_2) 2	6 4	3(5) 3	6 [total]	1(SO_2) 2
3 5 min	15 15 min	demand or 1 hour 1 hour	5–30 5 min–4 hours	5 —	1 1 hour
7[18]	8	12	10[30]	6	31
0.02	0.03	0.003	0.05	0.05	0.25
topography/large sources	topography/ industry/pop.	—	pop. density/ geography	—	diffusion model/ sources
digital/voice line	digital/TTY line	digital/wire + microwave	analog (PWM)/voice line	analog/voice line	analog (PFM)/phone (120 Hz)
IBM 1801	—	CPU	PDP-8	—	Philips P-9201
printout (reads red over standard)	printer [CRT]	real-time graphic, hourly printout	TTY printer	—	printer
COH—met. regression model	[2–24-hour forecast diffusion model]	analytical diffusion model	—	—	statistical (mean deviation)
—	[optimizing short- and long-term model]	—	—	—	—
meteorology forecast + SO_2/CO/ particulate levels	SO_2(0.30 ppm), 48-hour stagnation	O_3/CO/SO_2/NO_x	meteorology forecast + SO_2/CO/ particulate levels	—	meteorology forecast + SO_2
source abatement	fuel change or shutdown	3-stage alert (burning, traffic)	3-stage alert	—	source curtailment
[computerized]	large sources [computerized]	—	—	—	major sources known
SAROAD	—	—	SAROAD	—	NA
29, 30	31, 32	33	34–37	—	38, 39

range items, such as criterions or mathematical-model development, to help justify expensive data processors.

In measurement and sampling, there are two schools of thought. One utilizes SO_2 as a "tracer," or index of general pollution, whereas the other finds it necessary to monitor individual chemical variables. Those relying on SO_2 are either less concerned with automobile-derived photochemical smog or anticipate a strong correlation between pollution components. This difference is also reflected in the number of meteorological variables considered necessary.

The most striking inconsistency lies between the design of the sampling systems, particularly the station density, and the criterions for "representativeness" of the samples. Studies of SO_2 pollution in several areas of the world have agreed that the spacing between sampling stations should not exceed 0.8 km to obtain reasonably accurate estimates of the daily average. One standard, that of the West German Federal Republic, approaches this measure by requiring one station per square kilometer for pollution surveys. The actual densities tabulated for U.S. networks are at least two orders of magnitude below this criterion. Attempts have been made to justify this parsimony theoretically by appeal to diffusion models and by placing sensors near large sources. It has been shown in the Allegheny County system, for one, that this does not work.[30] In Rijnmond, however, where the topography and source parameters appear less complicated, it may prove possible to obtain accurate results with about one tenth the theoretical coverage. If this proves to be the case, it will serve as economic justification for greater application of diffusion modeling in support of sampling network design.

The data-handling systems should show few surprises,

since the current numbers of sensors do not require any novel techniques. In view of a debate on manual or analog versus digital data transmission, it is noteworthy that most networks have chosen or have switched to the digital technique.[33] The use of general-purpose computers for data processing is practically universal. In several systems, elaborate diffusion models are planned and, in at least one case, these are for short-term pollution forecasting and "feed-forward," real-time control of abatement procedures. But, to date, it has not been reported that any of these models have been implemented.

Conclusion

A cost-effective approach to air-pollution control, although by no means unanimously approved, seems inevitable over the long term. Optimizing the use of our air resource as a sink will require a higher order of pollution monitoring than has been planned to date, and an order-of-magnitude improvement in urban meteorology and forecasting. In turn, this means more effort in the development of sensors, mathematical models of pollution dispersion, and the design of urban systems. It is apparent that current network sampling systems will require much expansion and a period of intensive development and testing before the point is reached where air pollution can be predicted and averted. We have made a beginning in this laudable ambition, but like Alice's queen in the "Looking Glass" story, we must run as fast as we can just to stay in one place.

The author appreciates the encouragement and thought-provoking discussion of many concerned persons within the Honeywell organization, including Irving G. Young, Robert L. Wilson, James M. Lufkin, and James E. Myers, as well as the substantial support provided by Ethlyn Thomson, librarian, and Marie Williamson, secretary. This article will form part of a book on this subject by the author and Dr. Young that will be published by Wiley–Interscience in the fall of 1972.

REFERENCES

1. Stern, A. C., ed., *Air Pollution*, 2nd ed. New York: Academic, 1968.

2. Smith, M. E., "International symposium chemical reactions lower atmosphere," Advance Papers, Stanford Research Institute, San Francisco, Calif., 1969, pp. 273–286.

3. Seinfeld, J. H., "Mathematical models of air quality control regions," in *Development of Air Quality Standards*, A. Atkisson and R. S. Gaines, eds. Columbus, Ohio: Merrill, 1970.

4. Pasquill, F., *Atmospheric Diffusion*. Princeton, N.J.: Van Nostrand, 1962.

5. Carson, J. E., and Moses, H., *J. Air Pollution Control Assoc.*, vol. 19, pp. 862–866, Nov. 1969.

6. Pooler, F., *Internat'l J. Air Water Pollution*, vol. 4, no. 3/4, pp. 199–211, 1961.

7. Turner, D. B., *J. Appl. Meteorol.*, vol. 3, pp. 83–81, Feb. 1964.

8. Miller, M. E., and Holzworth, G. C., *J. Air Pollution Control Assoc.*, vol. 17, pp. 46–50, Jan. 1967.

9. Lamb, R. G., "An air pollution model of Los Angeles," M.S. Thesis, University of California, Los Angeles, 1968.

10. Wayne, L. G., in *Development of Air Quality Standards*, A. Atkisson and R. S. Gaines, eds. Columbus, Ohio: Merrill, 1970, p. 199.

11. "Air quality criteria for hydrocarbons," AP-64, U.S. Dept. of Health, Education, and Welfare, NAPCA, Mar. 1970, chap. 5.

12. Kohn, R. E., "Linear programming model for air pollution control: A pilot study of the St. Louis air shed," *J. Air Pollution Control Assoc.*, vol. 20, pp. 78–82, Feb. 1970.

13. Schwartz, S., and Siegel, G. B., "Models for and constraints on decision making," in *Development of Air Quality Standards*, A. Atkisson and R. S. Gaines, eds. Columbus, Ohio: Merrill, 1970, p. 40.

14. Stern, D. M., Honeywell Inc., personal communication.

15. Katz, E. L., and Morgan, T. R., "Analysis of requirements for air quality monitoring networks," presented at Air Pollution Assoc. Annual Meeting, St. Louis, Mo., June 14–18, 1970.

16. Bailey, S. J., "Control practice in the electric power industry," *Control Eng.*, vol. 18, pp. 42–44, Sept. 1971.

17. Gent, M. R., and Lamont, J. W., "Minimum-emission dispatch," *Proc. 7th Power Industry Computer Applications Conf.*, Boston, Mass., May 24–26, 1971.

18. Croke, E. J., and Roberts, J. J., "Air resource management and regional planning," *Bull. Atomic Scientists*, pp. 8–12, Feb. 1971.

19. "Environmental Quality," First Annual Report, Council on Environmental Quality, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., August 1970, pp. 83–88.

20. *The Federal Register*, Mar. 31, 1971 (36FR5931).

21. Degler, S. E., "State air pollution control laws" (rev. ed.), Bureau of National Affairs, Washington, D.C., 1970.

22. Yocum, G. E., *Chem. Eng.*, vol. 69, July 23, 1962.

23. Johnson, W. B., "Lidar applications in air pollution research and control," *J. Air Pollution Control Assoc.*, vol. 19, pp. 176–180, Mar. 1969.

24. Black, G., "Air quality and the systems approach," staff discussion paper 104, Program of Policy Studies in Science and Technology, NASA Grant NGL 09-010-030, George Washington Univ., Washington, D.C., July 1970, pp. 17–18, 33, 39.

25. Hunter, D. C., "The air quality monitoring program in New York State," presented at Air Pollution Control Assoc. meeting, New York City, June 22–26, 1969.

26. Brodivicz, B., Sussman, V., and Murdock, G., "Pennsylvania's computerized air monitoring system," *J. Air Pollution Control Assoc.*, vol. 19, pp. 484–489, July 1969.

27. Wolf, P. C., "Carbon monoxide measurement and monitoring in urban air," *Environ. Sci. Technol.*, vol. 5, pp. 212–217, Mar. 1971.

28. Wilkins, P. E., "Monitoring for compliance,' Monitor Labs, Inc., San Diego, Calif., 1970, pp. 10–11.

29. Bloom, B., Allegheny County Air Pollution Control Board, personal communication.

30. Stockton, E. L., "Experience with a computer oriented air monitoring program," *J. Air Pollution Control Assoc.*, vol. 20, pp. 456–460, July 1970.

31. Stanley, W. J., "A real time air pollution monitoring program, Dept. of Air Pollution Control, Chicago, Ill.

32. Cramer, H. E., "Meterological instrumentation for air pollution applications," in *Environmental Pollution Instrumentation*, R. L. Chapman, ed. Pittsburgh: Instrument Society of America, 1969, pp. 15–16.

33. Mills, J., "Continuous monitoring," *Chem. Eng.*, vol. 77, pp. 217–220, Apr. 27, 1970.

34. Heller, A. N., and Ferrand, E. F., "The Aerometric network of the City of New York," Environmental Protection Administration, Dept. of Air Resources, New York, N.Y.

35. Klein, S., "New York City steps up war on foul air," *Machine Design*, p. 38, Dec. 19, 1968.

36. *The New York Times*, May 5, 1970.

37. Eisenbud, M., "Environmental protection in the City of New York, *Science*, vol. 70, pp. 706–712, Nov. 13, 1970.

38. Cabot, F., "So goes SO_2," *Ind. Res.*, pp. 70–72, Sept. 1970.

39. "A new approach to the prediction and control of air pollution," Philips Gloeilampenfabrieken, Eindhoven, Netherlands, 1960.

BIBLIOGRAPHY

Bibbero, R. J., and Young, I. G., *Systems Approach to Air Pollution Control*. New York: Wiley-Interscience (to be published).

Hamburg, F. C., "Some basic considerations in the design of an air-pollution monitoring system," *J. Air Pollution Control Assoc.*, vol. 21, pp. 609–613, Oct. 1971.

Machol, R. E., ed., *System Engineering Handbook*. New York: McGraw-Hill, 1965.

Singer, S. F., ed., *Global Effects of Environmental Pollution*. New York: Springer-Verlag, 1970.

Stern, A. C., ed., *Proc. Symp. on Multiple-Source Urban Diffusion Models*. U.S. Environmental Protection Agency, Air Pollution Control Office, Superintendent of Documents, GPO, Washington, D.C., 1970.

Economic conditions in the U.S. electrical, electronics, and related industries: an assessment

In the decade ahead, engineering opportunities will increase, but at a slower rate, with supply of and demand for electrical engineers generally in balance. Recommendations are made for U.S. Government action in several areas to improve the engineering climate

William O. Fleckenstein

*Chairman, IEEE Ad Hoc Committee To Assess U.S. Economic Conditions in the Electrical, Electronics, and Related Industries**

Derived from a report of the Ad Hoc Committee To Assess U.S. Economic Conditions in the Electrical, Electronics, and Related Industries, this article reports on the group's principal findings for the 1970s. Their study sees a slowing of the demand for engineers with the rate increasing at about 2 percent a year, with supply and demand for electrical engineers reasonably in balance Government spending in domestic areas will probably not offset military and space decreases until at least 1980. The electrical/electronics industry will show an average annual growth of about 7.5 to 8 percent per year. Government action to provide a better data base on manpower and industrial output, reduce trade barriers, stabilize the economy, and foster research and development through tax incentives is urged.

A select committee, appointed by President James H. Mulligan, Jr., has quickly assembled for the members of IEEE a summary assessment of present and future conditions affecting the economy. Contained in the eight-point summary are recommendations that would require government action including establishment of a more useful statistical base for future studies of this kind.

Background material supporting the summary is contained in the pages that follow. The charge to the committee and a list of its members are contained in the editorial box on the following page.

In summary, the deliberations of the committee have yielded the following views on the decade of the 1970s:

1. The demand for engineers will be increasing at a rate of about 2 percent per year, a slower pace than in past years.
2. Supply and demand of electrical engineers will be reasonably in balance, perhaps with some shortages.
3. Government spending in the domestic areas will not offset decreases that have occurred or are expected in military and space programs until late in the decade, if then.
4. The electrical/electronics industry will show a 7.5 to 8.0 percent average annual growth, with some areas showing growth substantially above this amount.
5. There is a need to reduce the number of individual sources of data on manpower and characteristics of the industry, and to establish a more comprehensive and reliable data base to serve a multiplicity of users.
6. There is a need for aggressive governmental action to reduce trade barriers so that competition can be as open and fair as possible.
7. There is an enormous demand for capital that can best be met by a stable economy and incentives to investment.
8. To maintain our technical leadership, more support for research and development, probably in the form of tax incentives, is needed. A level of R&D expenditures in relation to Gross National Product (GNP) at least com-

* This committee report is published as a service to IEEE members. As noted in the article, the views expressed here are those of the individuals comprising the Committee and not necessarily those of the Institute or of the organizations they represent.

Genesis and background of the report

Early in 1971, Dr. J. H. Mulligan, Jr., President of IEEE, requested the author to organize a committee to assess "economic conditions in the U.S. electronics, electrical, and related industries." The objectives of the committee were to review available historical data in the industry, assess the principal forces of change that are or will be acting, and make judgments as to what trends are likely in the five- to ten-year period ahead. The study was to focus on those segments of the industry that are of key interest to IEEE members. The assessment was intended to be broad in scope, and quickly available. It relies on the judgments of knowledgeable people, rather than on an in-depth, time-consuming study.

With this charge in mind, an *ad hoc* committee was established whose members have a wide range of backgrounds that include technology, marketing, finance, economics, and various industrial interests. The members are: V. J. Adduci, president, Electronic Industries Association; E. Q. Daddario, senior vice president, Gulf and Western Precision Engineering Company; D. E. Eckdahl, vice president, manufacturing operations, National Cash Register Company; W. O. Fleckenstein (Chairman), executive director, Switching Systems Engineering Division, Bell Telephone Laboratories, Inc.; T. W. Folger, vice president of research, Kidder, Peabody and Company; W. F. Glavin, group vice president, Xerox Corporation, and president, Xerox Data Systems; D. L. Grove, vice president and chief economist, IBM Corporation; C. L. Hogan, president and chief executive officer, Fairchild Camera and Instrument Corporation; A. R. McCord, group vice president, Texas Instruments Inc.; J. M. Kinn (Secretary), Director, IEEE Educational Services.

The committee hoped initially that sufficient historical data might exist to allow a breakdown of the industry into major segments, probably by Standard Industrial Classification (SIC), and a compilation of significant factors for each segment. These factors might include sales, employment of engineers and scientists, capital expenditures, imports, and exports. From such data, forecasts based on anticipated forces and trends were planned that would summarize the committee's judgment on the future of the industry.

Numerous sources of data were used, including publications of Electronic Industries Association (EIA), National Electrical Manufacturers Association (NEMA), National Science Foundation (NSF), Department of Labor, Department of Commerce, and many others. In addition, representatives of a number of other industry organizations, including Aerospace Industries Association (AIA), Business Equipment Manufacturers Association (BEMA), and Scientific Apparatus Manufacturers Association (SAMA), were consulted. Each of these organizations compiles information appropriate to its own purpose, but does not endeavor to compile all data within a given industrial category.

The committee concluded that there exists no comprehensive data base for the electrical/electronics industry that is reasonably complete and nonoverlapping. Specialized segments exist, for example, EIA and NEMA, but there is no feasible means for making a "generation breakdown" that would allow recombining the data into a broad characterization of the industry. Although volumes of data are available from such sources as the U.S. Department of Labor and the U.S. Department of Commerce, the data do not exist in a form that is usable for a study of this kind. This problem will be encountered by any major scientific or engineering group attempting to make similar projections.

The lack of reliable data was a major concern to the committee and it is recommended that steps be taken as quickly as possible to explore the possibility of having a centralized government agency take the lead in establishing requirements for a data base that would be useful to professional societies, educational institutions, and industrial associations as well as the Government itself.

Faced with these limitations in source data, the committee proceeded on an important but more limited course. Estimates were made of engineering employment, including electrical engineering employment, of growth rates in certain segments of the industry, particularly those most heavily represented by IEEE membership, and of effects of other significant forces such as government spending, international trade, and capital investments.

parable to that of the mid-1960s would be more appropriate.

The results summarized in the foregoing derive from information that is grouped in the following categories covering engineering employment, government expenditures, growth in the industry, international considerations, and financial considerations.

Engineering employment

A significant result of the committee work is a forecast of engineering employment through 1975. Frequently, such forecasts are made by applying to an economic forecast some historical relationship between engineering employment and an economic factor. We found, however, that the direct relationship of engineering employment to economic aggregates—for example, gross economic indicators—was largely determined by trend. That is, both engineering employment and such economic variables as real GNP show strong upward trends over a long period of time. Therefore, they show a high correlation—but so would engineering employment be highly correlated with any time series that had a strong trend component. In fact, if the two series are related in some causal fashion so that knowing one you could predict the other, then the

year-to-year percentage changes would also be highly correlated. However, the analysis showed that there was practically no relationship between year-to-year percentage changes in engineering employment and in real GNP. For that matter, unlike variable growth rates in GNP, the growth rates for engineering employment exhibit a strong negative trend, as summarized by five-year growth rates in Table I.

The analysis also tested a variety of other economic indicators and a number of leads and lags. The final conclusion was that the direct relationship of engineering employment to economic aggregates is overwhelmingly dominated by trend, and, therefore, is not useful for making any discriminating forecasts.

Using another approach, we related real output to engineering employment for each of a number of industry groups. Measures of real output and engineering employment data[1] were available for the 1950–66 period for the following industry groups: durable manufacturing; nondurable manufacturing; construction; mining; transportation, communications, and utilities; wholesale and retail trade, finance, insurance, real estate, services, and all other private sectors; and government. The sum of these seven groups accounts for the total output and employment. By regression techniques we were able to develop a forecasting equation for the ratio of engineers to output in each of the industry groups. This ratio varies from industry to industry, and within an industry it varies as a function of time. Given a forecast output for an industry group, one could estimate engineering employment. The forecast of total engineering employment is simply the summation of the forecasts by group. The method was tested for the period 1967–70, for which period data were not available on engineering employment by industry group. The sum of the forecasts by group agreed reasonably well with the total number of engineers believed to be employed in these years. To provide the best possible base for the forecasts, a level adjustment was applied to each 1970 industry employment estimate, so that their total coincided with the actual total engineering employment for that year.

Forecasts of industry output for the years 1971–1975 were developed, using a sophisticated input/output model that, in turn, depended on a forecast of the economy. These estimates of output by industry sector, together with the regression equations relating engineering employment to output, made it possible to forecast engineers employed by industry group and in total. The resulting demand for engineers is shown in the upper portion of Fig. 1. Note that this forecast is designed to capture not only the impact of the economy, but also the changing industry composition of the total output of the economy and the changing productivity of engineers within each industry group. The forecasted distribution of engineers by industry category is shown in Table II.

The extrapolation of the forecast to 1980 was made by a different and much simpler method. The compound growth rate that resulted from the rather complex method described was used to extend the numbers to 1980.

Gross statistics of this kind do not pick up "minor" perturbations in total employment that can cause significant dislocations for a small percentage of the population. Thus, we did not see a dip in employment of engineers around 1970. In addition, since the supply of engineers is gradually increasing, this factor tends to mask the dip.

It should be emphasized that the forecasts derived here are based on the referenced historical data[1] and are neces-

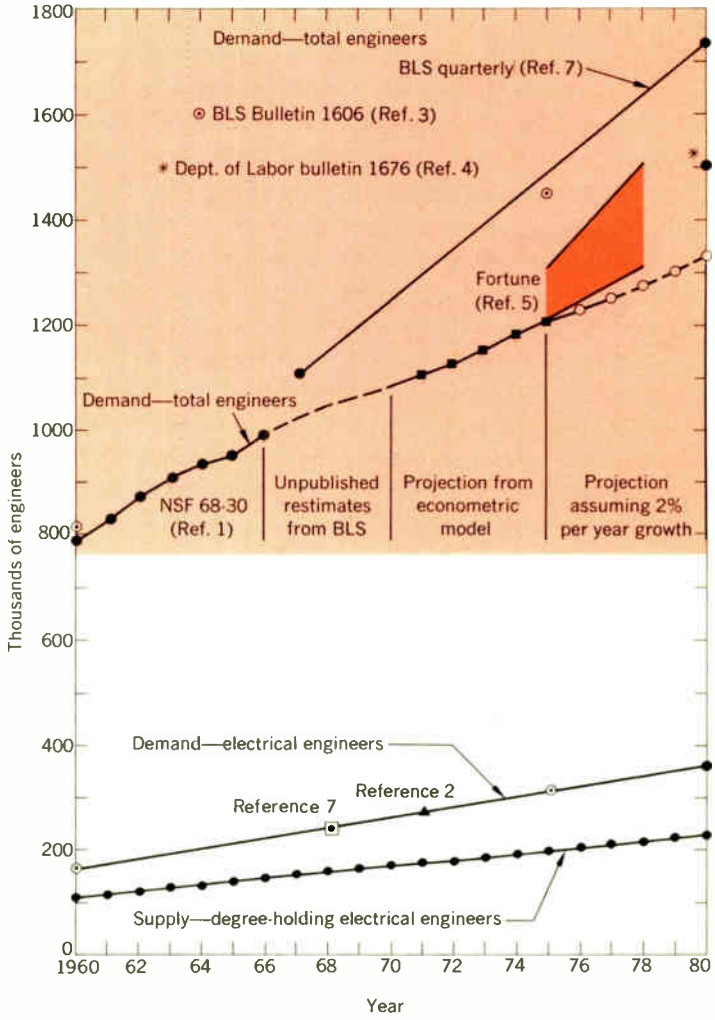
I. Growth rates of engineering employment and GNP in percent per year

	1950–55	1955–60	1960–65	1965–70
Employed engineers	8.1	5.9	3.9	2.4
Real GNP	4.3	2.2	4.8	3.2

II. Distribution of engineers by industry category, percent

	1955	1965	1975
Nondurable manufacturing	9.1	7.7	7.6
Durable manufacturing	42.9	46.3	52.0
Services	15.8	15.3	14.6
Construction	6.2	4.8	4.0
Mining	2.3	1.8	1.6
Transportation, communications	6.2	5.4	5.2
Government	15.3	14.8	15.0

FIGURE 1. Engineering employment, showing sources of data.



sarily dependent on the accuracy of these data.

It is useful to compare the demand curve derived here for all engineers with results presented by others. Various results obtained by others[1,7] are shown on the upper portion of Fig. 1. These data suggest that our forecasts may be conservative.

Data for electrical engineers are much more difficult to obtain. The demand curve for electrical engineers shown in Fig. 1 was plotted using available data that appeared to be mutually consistent. These data include members in the labor force who are characterized by the Bureau of Labor Statistics (BLS) and others as "electrical engineers," although many of them are not degree holders.

A question naturally arises as to our judgment in accepting these data on electrical engineers as reasonable since our derived demand curve for all engineers is lower than data from the same estimators. Our reasons are these: We believe that the relationship between our demand curve for all engineers and the demand curve shown for electrical engineers (showing a slightly increasing proportion of engineers being electrical) is likely to be correct. In addition, the source data for electrical engineers in some cases do not include closely related categories such as system analysts, programmers, and aerospace engineers that are categories included in the various estimates for "total engineers."

Still greater difficulty is encountered in attempting to determine demand for degree-holding electrical engineers. Indeed, there are no reliable data that the committee has seen indicating the *demand* for degree-holding electrical engineers. An estimate of the *supply* of degree-holding electrical engineers was arrived at in the following manner.

1. Information was obtained from the Bureau of Census indicating that there were 106 787 electrical engineers with four years or more of college education in the work force in 1960. (A reasonableness check of this number can be made by taking Department of Commerce figures quoted by John D. Alden,[2] for example, that 43 percent of engineers in 1960 were nondegree people, and assuming that the same ratio would apply to electrical engineers.)

2. Data on degrees granted, or estimated for the future, were obtained and are shown in Table III.[2,8] Although the discontinuity in B.S. degrees from 1970 to 1971 is unrealistic, and the advanced-degree data appear somewhat high, the error rate is believed to be negligible.

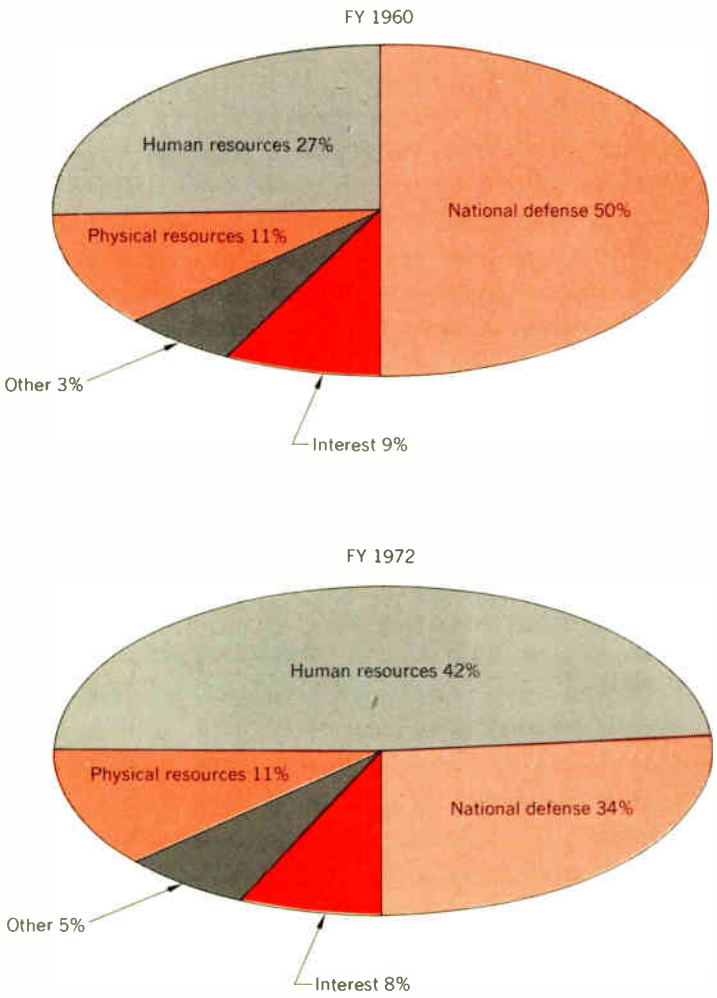
3. Starting with the figure 106 787 in 1960, the following formula was used to calculate the changes in the supply of degree-holding electrical engineers:

$$N_{t+1} = N_t - 0.024N_t + 0.9\,(BS_t - MS_{t+1}) + (MS_t - D_{t+3}) + 0.75D_t$$

where N_{t+1} = number of degree-holding electrical engineers in the work force in the year $t + 1$, N_t = number of degree-holding electrical engineers in the work force in the year t, BS_t = number of B.S. degrees granted in electrical engineering in the year t, MS_t = number of M.S. degrees granted in electrical engineering in the year t, MS_{t+1} = number of M.S. degrees granted in electrical engineering in the year $t + 1$, D_t = number of Ph.D. degrees granted in electrical engineering in the year t, and D_{t+3} = number of Ph.D. degrees granted in electrical engineering in the year $t + 3$. The following assumptions were included in deriving the results shown in Table III. It is assumed that 10 percent of bachelor-degree recipients who do not go on to graduate school in electrical engineering enter the labor force outside the electrical engineering classification; some are foreign students who return home. It takes one year to get the M.S. degree and three additional years to obtain the Ph.D. It is assumed that 100 percent of those obtaining M.S. degrees enter the work force, and 75 percent of those obtaining Ph.D. degrees leave the university and enter the regular work force. Of this work classification, 2.4 percent are lost each year owing to death, retirements, and transfers to other types of work (informal discussions with the Census Department suggest that this number is slightly low for electrical engineers). In line with point 5 of the summary, it is more specifically recommended that, in view of the difficulty in obtaining reliable statistical information covering the number and employment status of engineers (or electrical engineers), some action should be initiated by the IEEE or a group of professional societies to fill this need in the future.

The Joint Societies Employment Advisory Committee (JSEAC) might be an appropriate body to stimulate such an activity. It is our understanding that the National Science Foundation has funded related studies in the past.

FIGURE 2. Changing composition of the federal budget. The source is "The U.S. Budget in Brief, FY 1972."



Perhaps coordination between the professional societies and the NSF could lead to the development of a more accurate and comprehensive data base for the future.

Government expenditures

The shift in government priorities from defense and space programs to domestic programs is well known. The magnitude of this shift is illustrated in Fig. 2, which shows the composition of the federal budget in fiscal years 1960 and 1972, and in Table IV, which shows in somewhat more detail the rank order of priorities based on federal expenditures as a function of time. This transition has not been smooth, nor is it expected to be in the years immediately ahead.

It is expected that defense expenditures will decrease for the foreseeable future, and that the space program will continue to remain at the present level for a few years, and then edge upward to approximately $5 billion as the shuttle program develops and the Mars and Venus programs are determined. There will be substantial increases in expenditures for transportation, housing, education, health services, environmental control, and law enforcement. However, we do not expect that in the near future the decrease in defense expenditures for electronics will be offset by increases in socially oriented programs. This is not owing to any fundamental lack of need for technology in domestic programs, but rather to the difficulty in defining clear-cut goals that will generate broad support. It also results, in part, from the fact that defense expenditures have a higher percentage of total workers in the engineering and science category than in other industrial segments.

The early training of engineers and much of their professional experience are focused on what might be termed "physical" technology. In many parts of our society technologists are looked to largely as a means of applying new physical technology to solve specific problems and improve productivity. This view of the role of engineers has often led to the restriction of the application of technology to limited classes of problems.

To accomplish the transition that is being sought to solve socially oriented problems, it will be necessary for engineers to expand their understanding of society's needs to the point where they can participate in the optimum allocation of resources to meet those needs and contribute to meeting them through operational as well as physical technology.

We see this changing role of the technologist as essential but one that will come somewhat slowly and require his substantial education in the needs and operational mechanisms of our society. Therefore, we cannot foresee significant growth in technical manpower applied to socially oriented programs that are government funded until late in the decade.

With respect to public policy, the committee believes that the falloff in R&D as a percentage of GNP is an undesirable trend. Rather, a government policy that recognizes that the continuous development of new technology is an essential ingredient to the long-term health of the economy is most important. Such emphasis is required not only to provide the increasing productivity that is essential to the domestic economy, but also to maintain a suitable trade balance. The committee agreed that ways need to be found to provide economic stimulus and support to private industry to invest in R&D as the major means for providing the long-term growth required to maintain economic health. Such support might be pro-

III. Number of electrical engineering degrees granted by level and year

Year	Bachelor	Master	Doctor
1960	10 631	1 993	203
1961	10 200	2 400	250
1962	9 745	2 701	295
1963	10 393	2 816	386
1964	11 261	3 163	460
1965	11 670	3 566	511
1966	11 007	3 872	569
1967	10 843	3 942	688
1968	10 451	4 125	677
1969	11 375	4 019	851
1970	11 921	4 150	873
1971	10 200	4 340	870
1972	11 400	5 160	959
1973	11 700	5 640	1 070
1974	12 120	6 150	1 182
1975	12 500	6 630	1 293
1976	12 900	7 115	1 405
1977	13 240	7 625	1 517
1978	13 580	8 130	1 650
1979	13 930	8 600	1 750
1980	14 280	9 070	1 850

Source: 1960–1970, Ref. 2; 1971–1978, Ref. 8, plus the assumption that the proportion of EE's in future years will average the same as in past years; 1979–1980, estimated.

IV. National priorities as reflected in budget expenditures

FY 1960		FY 1965	
Defense	49.8%	Defense	41.9%
Income security	19.7	Income security	21.7
Veterans	5.9	Commerce and transportation	6.2
Commerce and transportation	5.2	Veterans	4.8
Agriculture	3.6	Space*	4.3
International aid	3.3	Agriculture	4.1
Natural resources	1.1	International aid	3.7
Housing*	1.1	Education and manpower*	1.9
Education and manpower*	1.1	Natural resources	1.7
Health*	0.8	Health*	1.5
Space*	0.4	Housing*	0.2
All other	8.0	All other	8.0
FY 1970		**FY 1972**	
Defense	40.8%	Defense	33.8%
Income security	22.3	Income security	26.5
Health*	6.6	Health*	7.0
Commerce and transportation	4.7	Commerce and transportation	4.8
Veterans	4.4	Veterans	4.6
Education and manpower*	3.7	Education and manpower*	3.8
Agriculture	3.2	Agriculture	2.5
Space*	1.9	Housing*	2.0
International aid	1.8	Natural resources	1.9
Housing*	1.6	International aid	1.8
Natural resources	1.3	Space*	1.4
All other	7.7	All other	9.9

* Identifies areas of significant change.
Source: "The U.S. Budget in Brief, FY 1972," p. 29.

V. IEEE U.S. membership distribution by SIC code, with forecast of growth rate in dollar value of shipments

SIC Code		IEEE U.S. Membership	1970–1980 Growth Rate, percent per year
36 Electric machinery equipment and supplies		47 310	
3662 Radio and television transmitting, signaling, detection equipment and apparatus	17 740		7.0
3679 Electronic components and accessories (not elsewhere classified)	5 960		
3612 Power, distribution, and specialty transformers	3 030		6.0
3611 Electric measuring instruments and test equipment	2 880		6.7
3651 Radio and television receiving sets, except communications type	2 510		6.0
3661 Telephone and telegraph apparatus	2 460		10.0
All others	12 730		
35 Machinery, except electric		11 700	
3573 Electronic computing equipment	7 180		10.5
3511 Steam engines, turbines, and turbine-generator sets	975		7.5
3572 Typewriters	515		3.9
3574 Calculating and accounting machines, except electronic computing equipment	315		5.5
All others	2 715		
49 Electrical, gas, and sanitary services		11 040	
4911 Electrical companies and systems	8 890		
4931 Electrical and other services combined (electrical less than 95 percent total)	1 880		6.0
All others	270		
37 Transportation equipment		9 330	
3721 Aircraft	5 880		—3.7
3729 Aircraft parts and auxiliary equipment (not elsewhere classified)	1 100		—0.3
3714 Motor vehicle parts and accessories	765		
3722 Aircraft engines and engine parts	625		—1.1
All others	960		
73 Miscellaneous business services		6 510	
7391 Commercial R&D labs	4 690		
7392 Business, management, administrative, and consulting services	940		
7399 Business services (not elsewhere classified)	730		
All others	150		
38 Manufacturers of instruments—professional; scientific, controlling; photo and optical goods; watches and clocks		6 170	
3811 Engineering, lab, scientific, and research instruments and associated equipment	2 870		6.7
3821 Instruments for measuring, controlling, and indicating physical characteristics	1 120		5.4
3861 Photo equipment and supplies	880		
3822 Automatic temperature controls	710		5.6
All others	590		
48 Communications companies		5 360	
4811 Telephone	3 750		9.0
4832 Radiobroadcasting	470		7.1
4833 Television broadcasting	455		9.0
All others	685		
Miscellaneous categories	27 580	27 580	
Total	125 000	125 000	
Composite growth rate (weighted)			6.1

Source: Membership estimated from 50 percent sample; value of shipments estimated from Ref. 9.

vided in the form of tax incentives. In addition, the provision of federal grants to industries whose R&D activities show promise in the solution of domestic social problems would seem appropriate. The latter activity might be modeled after the Canadian policy that allows that government to share 50 percent of the cost of such private R&D.

Growth in the industry

As noted, a suitable and comprehensive data base sufficient to forecast industry growth with reasonable accuracy does not exist. Consequently, we shall look at major segments of the industry of particular interest to IEEE members, indicate judgments on particularly strong areas of growth, and suggest forces that are operating and are therefore likely to create emphasis in certain areas of technology. A number of sources of information[9–13] were used; the estimates reflect the judgments of the committee.

Table V shows the distribution of IEEE membership in the United States by Standard Industrial Classification (SIC) code based on a 50 percent membership sample.

VI. Industry forecasts, 1970–1980

Group	Percent Increase per Year
Computers, computer peripherals, and memories	11
Communications (except radio and television)	9
Process and industrial controls	12
Test, measuring, scientific, and medical instruments	11
Transportation electronics (nonmilitary)	12
Semiconductor industry	9

Also indicated for certain sectors are growth rates in dollar value of shipments obtained from a U.S. Department of Commerce publication.[9] The growth rates obtained cover segments of the industry that represent an estimated 43 percent of IEEE U.S. membership. Based on these projections, a weighted growth rate of 6.1 percent is obtained. The forecast for electronic manufacturing sectors is higher and the committee's judgment is that an overall annual average growth rate for the electrical/electronics industry of between 7.5 and 8.0 percent is more likely over the next decade.

The combination of forces at work suggests that certain segments of the industry will grow more rapidly than the overall growth rate indicates. Among the more important of these forces are those being created by changes in labor, social values, and technologies.

The labor-related forces include unavailability of people to perform menial labor, welfare availability that competes with low-level jobs, minimum-wage requirements, and an apparent long-term trend of labor costs increasing more rapidly than other components of cost. All of these forces suggest the need for a more rapid rate of technical innovation and a more rapid rate of increase in capital expenditures in the future.

Social forces include environmental control, on which companies are beginning to spend substantial sums of money; the large body of law building up that will require a more vigorous response by industry; and the increasing application of technology needed to achieve the desired results. Improvements required in medical care, education, transportation, and law enforcement are examples of other social areas demanding increasing use of technology.

In technology, the rapid change in materials and device technology is leading toward components and systems for which costs per function are continuing a downward trend relative to other costs.

An analysis of these forces suggests that there will be a strong movement to automate the operational aspects of businesses in the decade ahead. In addition, instrumentation and process control will see heavy growth. Table VI indicates areas in which the committee believes strong growth rates will be encountered during this period. The arguments put forth lead almost directly to those disciplines that we believe are likely to receive special emphasis in the next decade.

Device technology is continuing toward more complex operations in which a large number of batch-processing operations are carried out in sequence. This trend will continue to place strong emphasis on materials and process technology, including physics and chemistry, materials engineering, mechanical design, and electrical design. Integrated semiconductor technology will see heavy growth in both linear and nonlinear applications, and particularly in memory applications. Optoelectronics is also an area in which significant growth can be expected.

Computer-aided design, which has received considerable emphasis in the past decade, will become much stronger in the decade ahead. This technique not only includes analysis, but, more important, the portions of the "design" job that can be sufficiently formalized to automate. In addition, simulation of the performance of systems and subsystems will become increasingly important.

Sensing and control instrumentation, all aspects of computer technology, and communications—particularly digital communications—are key aspects of almost all of the growth areas noted.

International considerations

Competition from outside the U.S. has been an increasing problem for U.S. industry in recent years. Principal reasons are an exchange-rate structure that has overpriced the dollar, lower labor rates abroad that do not reflect lower productivity, changing technological conditions, and government policies. The following will cover broad changes in technological development and their implications, and some recommendations for possible governmental action. It should be noted that this review, including the section that follows, was carried out prior to the recent establishment of economic controls by President Nixon.

With regard to current and continuing trends in the area of foreign trade, and more specifically in the matter of U.S. competitiveness with foreign industry, two principal technologically oriented factors stand out. The first, and perhaps the most fundamental trend, is that concerned with the size of the technological gap between the U.S. and other technically strong countries such as Japan and Germany. There is little question that the technological growth rates of these and other countries have exceeded that of the U.S. in the recent past. It is likely that the magnitude of the "gap" will continually decrease, although at a gradually diminishing rate. This narrowing of the technological gap between the U.S. and the other countries is likely to be enhanced by the recent withdrawal of federal R&D support that has so significantly sustained and stimulated the U.S. level of technological growth in the past.

The second factor is a consequence of general advancement in technology and rapidly increasing labor costs; both are causing industry to move increasingly toward products involving low labor content but high material and technology content. This combination, coupled with the expanding and deepening technical capability across the international spectrum, is changing the economic tradeoffs that multinational companies use in making geographical determinations on product design as well as manufacture.

As an expected consequence of these trends, it is highly probable that increasing use will be made of international technical capability by U.S. multinational industry, on the presumption that the ultimate market location will establish not only the site for product *manu-*

facture but also the site for product *design*. Undoubtedly, although more than counterbalanced by other influences, this particular trend suggests a negative influence on engineering employment prospects within the U.S.

The most desirable and probably the most effective approach for broadening U.S. technical employment opportunities, insofar as they may be influenced by international trade factors, involves developing the maximum amount of trade between nations. Under such conditions, where head-to-head competition is a reality, each country's technological skills of all kinds—research, engineering, and manufacturing—are challenged, provided there is reasonable parity of capability, a condition increasingly coming into being. In meeting such a challenge, the technologies involved are strengthened and employment opportunities increased.

To achieve this enhanced international interchange, an appropriate U.S. Government trade posture is essential. Several important aspects of that posture include elimination of reduction of quotas, governmental support for exporters, reduction of trade barriers, and relaxation of controls on foreign direct investment.

Quotas are not the answer to the declining trade surplus or to the question of import competition. Rather, they lead to still higher costs and prices that freer trade and greater competition would serve to combat. International competition not only offers consumers access to inexpensive imports, but it restrains price increases in the domestic industries that face such competition. Moreover, the imposition of quotas can only invite retaliatory measures by other countries.

Efforts to expand U.S. exports could be assisted by a U.S. policy that would offer exporters the kind of financial assistance and tax treatment now used by some other governments to support their exporters. A current proposal for a Domestic International Sales Corporation (DISC) under consideration by the U.S. Government is commendable in this connection.

Tariff and nontariff barriers around the world continue to be a significant impediment to U.S. export expansion. However, nontariff barriers are one of the most critical and difficult problems facing U.S. trade policy. Both tariff and nontariff barriers should come under aggressive attack by U.S. trade negotiators. It would seem difficult to assume this posture in an environment of restrictive trade at home.

The overall international trade questions cannot be viewed in isolation from the other components of U.S. foreign economic policy. This is particularly true of the foreign investment question that has assumed steadily increasing importance during the last decade. Direct U.S. investment has become larger and is growing faster than exports. In the decade ahead, it is important that this growth continue and that it be supported by U.S. Government policy, as U.S. investors will have to compete with increasingly aggressive investors from Japan and Western Europe for the substantial returns that can be realized. The U.S. Government's preference for controls on foreign direct investment to meet short-term balance-of-payments goals should be examined in light of the fact that foreign investment is a major long-range source of foreign receipts.

In view of the need for tying the numerous threads of U.S. foreign economic relations into a unified policy, the President's action to establish the Council on International Economic Policy as a new mechanism that will plan and coordinate all U.S. foreign economic policy is appropriate. This step will give foreign economic policy the priority and cohesiveness essential to achieving the most advantageous foreign economic posture in the 1970s.

Financial considerations

In the 1970s, as in the 1950–70 period, trends in the securities markets will have an effect upon the employment of technically trained personnel in the United States. In the 1950–70 period, many new technological enterprises were started with a generous flow of funds from the financial markets, including a plentiful supply of venture capital. Since World War II, the financial markets have looked kindly upon new technological enterprises. In addition, established firms have found the money markets willing to supply funds for new technological ventures. Also affecting the employment of engineers are the expected levels of capital spending by all industry, since engineering manpower is required to design new plants and modernize existing facilities. Levels of capital spending are closely related to the trends of corporate profits and are dependent on many other factors; among the most important are the *availability* and *price* of money to fuel capital expenditures.

Examination of the sources and uses of corporate funds since 1960 shows that since 1966 there has been a rapidly increasing reliance by corporations on external sources to finance plant equipment expenditures and to meet other financing needs. The primary internal sources of corporate capital—undistributed corporate profits and depreciation—have not kept pace with corporate needs for funds. Expressed as a percentage of Gross National Product, corporate profits after taxes fell precipitously in the late 1960s to a low in 1970 and 1971 not seen since the days of the Great Depression. Depreciation as a percentage of GNP rose during the 1960s, but the combination of corporate profits and depreciation has been declining as a percentage of total GNP and is currently around the level of the recession year of 1960. However, this relationship should improve as the economy recovers from the recent recession.

On the other hand, corporate expenditures on plant and equipment rose dramatically both in absolute amount and as a percentage of GNP during most of the 1960s, falling off only in the 1969–71 recession. There are reasons to believe that capital investments should remain high during the 1970s. The United States must modernize and automate its industries if it is to offset rapidly rising labor costs and remain competitive in the international arena. To combat the pollution of our environment, enormous capital expenditures are going to be required by many industries. The inflationary trends alone dictate a tendency toward higher capital expenditures because the same dollar a few years later will not buy the same amount of actual physical plant and equipment. The replacement cost of plant and equipment is continuously rising. In general, it can be said that inflation is a great consumer of capital. As our society becomes more technologically complex, the amount of capital investment per employee will continue to rise.

The other side of the coin is the supply of capital and the price at which it will be supplied. Studies have shown that long-term interest rates are closely related to ex-

pected inflation rates. An approximate formula for predicting long-term interest rates in an inflationary environment is that each percentage point of expected inflation rate adds a percentage point to long-term interest rates. We cannot predict what interest rates will be, but, in general, we do believe that interest rates will be higher in the early 1970s than they were during most of the 1960s, owing to expected higher rates of inflation. Both factors—inflation and strong demand for funds—will be operating in the early 1970s to produce high interest rates.

Because of the increasing reliance of corporations on external financing, we also see that the financial markets will be called upon for more capital in the 1970s than during most of the 1960s. We do not predict a capital shortage that would cause any prime customer or major solvent corporation to be unable to obtain sufficient funds from the financial markets, but we doubt that the new enterprise or the small technological company can count on being as readily received and financed in the 1970s as it was in the 1960s, except where the expected growth rate of earnings is very promising. Also, owing to the higher price that we foresee capital will command in the 1970s, a number of capital improvement projects may not prove as economical as they might if undertaken in the 1950s and 1960s. Although the level of capital spending will be high in the 1970s, we conclude that the financial markets may exert a downward influence upon the level of capital spending.

With regard to basic financial considerations, insofar as they affect electrical/electronics engineers and the industry, the committee made four conclusions:

1. The most obvious and greatest assistance to the financial markets would be the restoration of an atmosphere of stability with only moderate inflation in the U.S. economy—much less than the current 5 to 7 percent annual rate. There is nothing like a stable dollar to foster stable financial markets. The allocation of the country's resources can be much more wisely and fairly distributed without the conditions of worrisome inflation.

2. The United States must provide for more realistic replacement of our worn-out assets through a more realistic depreciation program. Current production must be charged with costs of replacement and modernization of its production equipment, and we are only inviting disaster unless this condition is met. The recent liberalization of depreciation rates is a commendable step; we recommend a continuing review of depreciation schedules. In addition, tax programs such as an investment-tax credit would also provide incentive for capital projects and help increase the internal generation of funds to finance capital expenditures.

3. To allow the operation of stable financial markets, we must have a stable monetary policy. The severe swings of interest rates or of money supply serve only to disturb the financial markets.

4. As suggested in a previous section, if the United States really wants to exert every effort to maintain its technical leadership in the free world, it is recommended that consideration be given to using special tax incentives to encourage scientific research and development. Several alternatives are available, such as increased tax deductions for research and development by corporations and special tax treatment for individual capital gains and losses incurred in venture capital or small business investments.

In addition to the individual efforts of the committee members (listed in the box), support and counsel were received through them from colleagues in their various organizations. Without this assistance, much of the work would have been impossible.

REFERENCES

1. "Employment of scientists and engineers in U.S.," Bureau of Labor Statistics, NSF Bulletin 68-30.

2. Alden, J. D., "Manpower trends in electrical engineering in the United States," *Proc. IEEE*, vol. 59, pp. 834–838, June 1971.

3. "Tomorrow's manpower needs," Bulletin 1606, U.S. Dept. of Labor, Bureau of Labor Statistics.

4. "College educated workers, 1968–80," Bulletin 1676, U.S. Dept. of Labor.

5. Gooding, J., "The engineers are redesigning their own profession," *Fortune*, June 1971.

6. "The occupational outlook for engineers and technicians," *Engineering Manpower Highlights*, Engineers Joint Council, Nov. 1970.

7. "Quarterly occupational outlook—Summer 1970," Bureau of Labor Statistics.

8. Engineering Manpower Commission, U.S. Office of Education, and Commission on Human Resources and Advances Education, quoted in EMC Bulletin No. 17, Sept. 1970.

9. "U.S. industrial outlook 1971, with projections through 1980," U.S. Dept. of Commerce.

10. "The U.S. economy in 1980," U.S. Dept. of Labor.

11. "Projections of gross output by industry," National Planning Association.

12. "World electronic industries—new opportunities for growth and diversification in the 1970's," Stanford Research Institute (informal discussion of study with K. W. Taylor, SRI).

13. *Electronic Market Data Book 1971*, Electronic Industries Association.

Reprints of this article (No. X71-125) are available to readers. Please use the order form on page 8, which gives information and prices.

W. O. Fleckenstein (F) is executive director of the Switching Systems Engineering Division at Bell Telephone Laboratories, Holmdel, N.J. He has previously served the Bell System as general manager of R&D for Western Electric in Princeton, N.J., and as an executive director in the Bell Laboratories, responsible for development of data communication, private branch exchange, telegraph, key telephone, and private line systems. During his 19 years with Bell Laboratories, Mr. Fleckenstein has made important contributions to development of the first data sets and participated in the early development of electronic switching systems, in which area he has been granted a number of patents. He had been director of a group editing a four-volume series on "Physical Design of Electronic Equipment," now being published.

He majored in communications at Lehigh University, earning the bachelor of science degree in electrical engineering, with highest honors, in 1949, and winning the Dupont Memorial Prize. In 1959, he received honorable mention from Eta Kappa Nu as an "outstanding young electrical engineer."

Mr. Fleckenstein was chairman of the Technical Program Committee for the 1970 IEEE International Convention, and is a member of the Board of Directors and the Executive Committee of IEEE.

New product applications

Power package for small vehicles uses improved battery-charging system

Increasing use of electric propulsion for small vehicles has shown a need for more economical battery service. Typical applications include golf carts, riding mowers, garden tractors, and fork-lift trucks. In Fig. 1 a battery-charger unit is shown installed in a 1972 model riding mower.

FIGURE 1. Power package installed.

The typical vehicle power system shown in the block diagram of Fig. 2 encompasses all the necessary functions that must be included in any application, and some or all of the units at lower right will be required in more sophisticated vehicles. Heart of the power package is the recharging unit, shown outside its protective enclosure in Fig. 3.

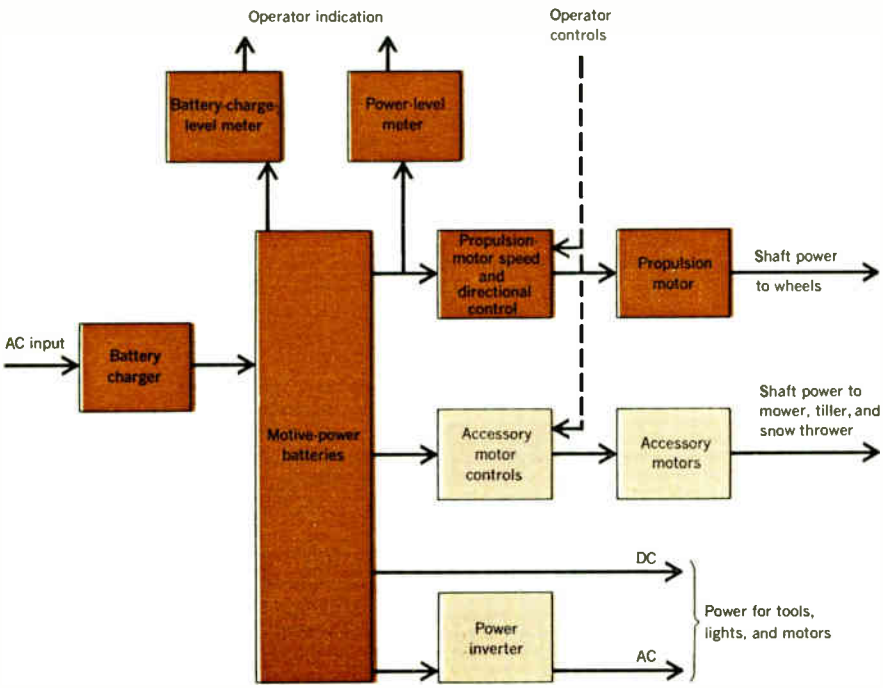
Typical charge-voltage-versus-current characteristics of a battery at the end of charge are shown in Fig. 4. The end-of-charge voltage is a strong function of battery age and temperature. Conventional charges show a constant-voltage characteristic, so that the final charge current is given by the intersection of the charger curve with the battery characteristic curve. For a new 80°F battery, final charge current is an acceptable 5 amperes. An old 120°F battery receives a final charge of 18 amperes that is excessive. Increased temperature, gassing, and water consumption reduce battery life. A new battery at 0°F receives a final charge current of only 0.4 ampere and may not be fully charged. Besides, a fixed charge time is usually set at the discretion of an operator, and may result in excessive undercharge or overcharge. The effect may be insufficient battery performance and potential battery sulfation. A means must also be provided to maintain charge during long storage periods.

The new Gould reactance-limited charger shows the voltage–current characteristic given in Fig. 4. The narrower final-charge-current range, from 4 to 11 amperes, reduces problems of over- and undercharging of batteries at temperature and age extremes. The Gould charger uses a compensated voltage-sensing circuit that automatically starts a timer when battery-charge voltage reaches a point when the battery is about 85 percent recharged. The timer provides a fixed-time end of charge. The start of the timer is automatically adjusted for battery temperature.

Write for details to Gould Inc., P.O. Box 3140, St. Paul, Minn. 55165.

Circle No. 55 on Reader Service Card

FIGURE 2. Essential functions of power package, and ancillary (lower right).



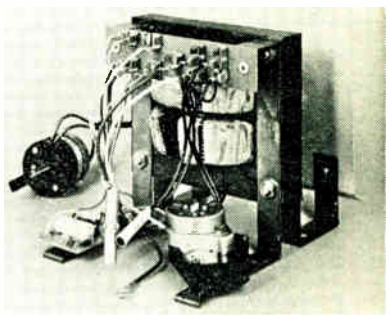


FIGURE 3. Elements of charger.

FIGURE 4. Battery-charger characteristics for various battery temperatures.