Journal of The Lepidopterists' Society

Volume 20 | 1966 | Number 1

APPLICATION OF AUTOMATION IN RHOPALOCERA RESEARCH

Charles R. Crowe
5027 N.E. 23rd Ave., Portland, Oregon

In the 1890 census of Baltimore, Maryland, Dr. Herman Hollerith introduced the first punched card method of tabulating data (Copeland, 1949). Since that time, punched card procedures have mushroomed to the point where automated tabulating processes are being utilized in almost all phases of science and industry.

To date, automated procedures in the Rhopalocera field are few. It is true that some experimental, and highly controversial, work has been carried on in the taxonomic area—witness the recent papers of Ehrlich, Rohlf, Sokal, Michener, and Sneath, among others. The scope of this pilot work is quite limited as of yet, though implications indicate a much broader investigation of this lucrative field by butterfly taxonomists.

From compilation of punchcard data, the volumes of *Forest Lepidoptera of Canada* (McGugan, 1958; Prentice, 1962, 1963) have been realized, and forest entomologists in the United States are also accumulating data on punchcard file systems (Powell, 1965).

My desire, stemming from a speculative note by J. W. Tilden in the March, 1962, issue of the *Lepidopterists' News*, is to acquaint the average collector with automated tabulating machines, with the primary emphasis on recording field records and observations of collectors, and standardization of collecting procedures.

The equipment mentioned is commonly in use today in banks, offices, and department stores and is mentioned here due to its relatively low cost. Many more elaborate computing or tabulating systems, such as International Business Machines (hereafter abbreviated to IBM) highly versatile 1401 series (Anonymous, 1961), are on the market today.

As we frequently see in the news media, the Federal Government is underwriting huge sums annually to other deserving aspects of natural science. Perhaps a grant from the National Science Foundation or a similar organization would be sufficient to start the project. Already Cornell

University has produced a *Pilot Register of Zoology* (W. L. Brown, 1964) for card listing of all living and fossil organisms. Perhaps soon a project of that type may be developed on a continuing basis.

The U. S. Department of Agriculture presently has in operation a system adapted to the study, by automatic data processing, of range plants (Garrison & Skovlin, 1960). The plants each have a number, in a general catalog, and when noted in the field, a coded report is sent to the central office where it is punched in a card and processed on tabulating machines.

Since 1950 the Canadian Forest Insect and Disease Survey has utilized standardized procedures of recording field data on punch cards. At present, close to a million records have accumulated, with an annual inflow of new records close to 50,000 (C. E. Brown, 1964). In 1962, programming a Univac Solid State Computer, maps of Lepidoptera distribution were produced for Canada, utilizing methods described by C. E. Brown (1964).

THE PUNCH CARD

The punch card, when properly prepared, can produce automatically, and at high speed, an almost unlimited number of statistical comparisons (Arkin & Colton, 1964). The card is the actual controlling agent of all the machines through which it passes, causing itself to be counted, printed, classified, sorted in a sequence, collated in sequence, compared, selected, reproduced, edited, coded, and decoded, plus doing all the normal arithmetical operations.

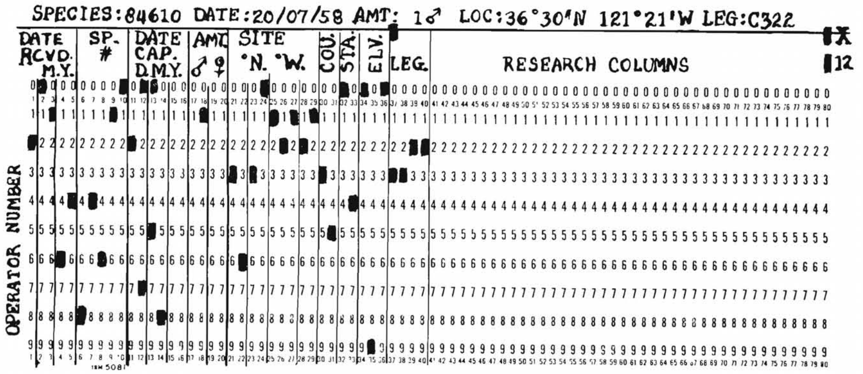
The card is divided into 12 horizontal columns and 80 vertical columns. The horizontal values indicate a general topic, as date or species, while the vertical values give a specific value to each subject. In Fig. 1 is reproduced a standard IBM punch card. The specific fields (areas to be punched, like date and species) may be marked off in any desired arrangement.

The vertical columns have 12 general positions for punching, the two at the top being known as 12 or X punch areas, which are used as controls, or to identify a specific card. Thus an X punch in column 47 might indicate that the card carries special data on *Euphydryas*.

Some possible card field descriptions, with regard to butterfly data, are as follows:

Date card punched. The first card field, or first five columns of the IBM card, would be reserved for the date the card was actually punched and a code number for the keypuncher.

Species number. Card columns six to 10 could be reserved for a number



EXPLANATION OF FIGURE

Fig. 1. The IBM punch card, as would conceivably be used in an entomological IBM installation.

to be ascribed to each "species." This number does not, of course, have to be five spaces long, but that gives sufficient room for general locality records of some 99,999 different varieties to be numbered.

Locality. Present methods of citing localities will not suffice for use on IBM cards. The only logical approach for a punch card is with latitude and longitude. This would necessitate collectors keeping track of this new information for each collecting site.

County and state. Vast numbers of cards would be sorted by species and geographically, to facilitate manual handling. This would be accomplished by coding the county and state of the locality from any standard atlas. The numbering would not need to be limited to North America, but could cover the world.

Elevation. The altitude of capture would best be stated in hundreds of feet or meters above sea level.

Collector. Each collector would have an assigned number, determined automatically as data is received from that collector.

Since the machines print alphabetical figures, the first figure of an individual's number would consist of the first letter of his last name, i.e., John Doe, being D422.

Remaining columns. Those columns remaining could be utilized in any means desired. Any type of information could be coded by number, the data being dependent on standardization—the key word in automation. The areas could be used in any field of research, with entries of taxonomic, climatic, genetic, or other interests. These columns may be X or 12 punched in any specific column to indicate any one of several hundred different areas of research. The specialist, then, need only develop a code for his area of study, and publish it in a suitable medium.

I emphasize the policy of nonspecialists doing the basic fieldwork specifically for the purpose, that it is intolerable that the specialist should have to do all his own sampling (Ehrlich, 1961b), often covering years of research, when abundant punch card data could be within easy reach.

Determinations. It is rare, even with experts, when 100% accuracy in determinations is realized. The use of numerical taxonomy should quantify this greatly, by giving numerical limits to degrees of variation. This data, if to be properly utilized, must then be presented in a usable form to novice collectors, so that accurate carded data may be forthcoming.

Perhaps then, general instructions could be presented to collectors, developing a special taxon, procedure, or covering a related area such as ecology, genetics, or taxonomy. After completion of these instructions, specialists may reasonably expect a high degree of accuracy in the information submitted concerning those specialized areas studied by the amateur. In turn, one can also expect the novice to become proficient, and eventually a specialist in his own area of interest. If a pilot project of this sort could be sponsored by the Society, it might be well to investigate it further.

THE DATA PROCESSING MACHINES

The actual data processing system consists of numerous machines, each designed for a specific purpose, and all with an extreme range of function.

The keypunch. Data received from collectors must be punched in code numbers onto cards. This is done on a keypunch, a typewriter-like machine that punches holes rather than printing. The holes are then read by various machines, through which the card passes, by means of metal brushes completing a circuit through the cards, the body of the card acting as an insulator.

The reproducer (IBM 514). Where large amounts of similar information are to be punched, a reproducer is utilized, which copies, or "gang punches," the data from the preceding card.

When connected to the IBM 407 tabulator, the 514 acts as a "summary punch," punching totals accumulated on the tabulator into new cards to reduce handling of large volumes of similar cards.

The interpreter (IBM 552). When data is to be manually handled a great deal, as will large volumes of butterfly record cards, the punched holes may be printed on the top of the card, so as to be read visually.

The sorter (Several makes). The sorter is used to put the cards into any type of sequence or order. This machine can select, group, or reject the numbers on any single column. Quite probably, new incoming cards will be sorted by collector, county, state, and species (sorting the most minor area first), and collated with the main file.

The collator (IBM 085). To facilitate filing of the cards, a collator is used to merge groups in a common sequence. New cards will then be easily merged with the main file, rather than a time-consuming sorting process, due to the 085's ability to read 80 columns simultaneously. This machine may also be used to select groups of common card fields or control punches.

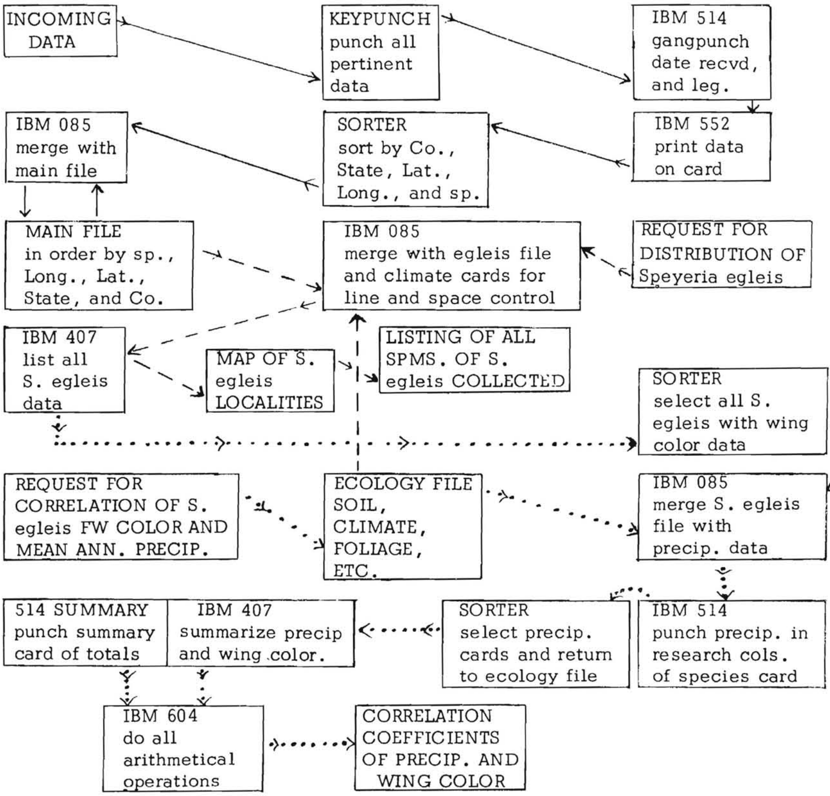
The tabulator (IBM 407). The tabulator is used to list all the coded data. It is another machine with many uses, and can add and subtract the card data while printing, besides cut a summary card on the IBM 514. A wiring panel on this machine gives an almost unlimited field of alternate working functions, each function being controlled by individual control wires on the panel. The IBM 514, 552, 085, and 602–604 systems also have this control panel.

The computer (IBM 602–604). Oftentimes there is a need for mathematical formulas to be solved in large numbers. This is done by a computer, and can handle most problems encountered in statistical comparisons, correlations, or ratios. A disadvantage of this specific machine is its relatively slow speed.

Operation of the Machines

Figure 2 shows the flow of information from collection of data to its dispersal in a typical operation. A quick sort yields all the cards of a species in a given region for those who desire listings of all available information on distribution in an area. By sorting on latitude and longitude, the IBM 407 can put a series of dots in sequence on a sheet of paper, the student needing only to add a map outline which has been printed to scale, in order to receive the known distribution of that species in the area of interest (see C. E. Brown, 1964). Another request on correlation of wing length and precipitation would be summary punched to find totals, the cards having sums figured on the computer. Thus, any type information desired, if punched on the cards, may be arranged in any manner and printed, whether a simple sort is required for a distributional study or a complex selection process comparing several factors against each other is desired.

In practice, quite probably collectors would be asked to supply the raw field data to be punched in the cards described. Figure 3 is a suggested format for consolidating field data. The area to the left would



EXPLANATION OF FIGURE

Fig. 2. The flow of data cards through an IBM installation. Solid line indicates path of new data to main file. Short dashes indicate path of data in request for distribution. Dotted line indicates path of data in request for correlation of data.

serve for the IBM card, the balance of written material being for the collectors' convenience. The format is not copyrighted, and may be copied.

In the checkered area at the bottom of the card, coded data of research interest would be written. The control figures on the left indicate what the following information might consist of, and how to control it. On the card, "Taxo. X-45" indicates taxonomic data, the control punch an X in column 45. The circled column indicates the first punch and last punch of a field, the number being the number of the column. The coded number itself will represent a value for a standard taxonomic feature, as genitalic measurement ratios or wing pattern elements.

CODED DATA	
05360:SPECIES 120764:DATE 1 male:AMOUNT 43°35':LAT. 119°56':LONG. 13 :COUNTY 37 :STATE 50 :ELV. C211 :LEG.	SPECIES: Lycaena cupreus Date:12 Jul 64 Co: Harney State: Oreg. SPM # 1256 LOCALITY DETAIL: 12 Miles NNW of Burns, Hy. 395. In small ravine in Pine forest, along Theimer Creek. TWP: 21S RAN: 31E SEC: center of 15 WEATHER: abt. 85°, few cumulus, wind light PLANTS NOTED: Willow and Aspen underbrush, Lodgepole Pine dominant, Lupines abundant. Leg:C.R. Crowe Det by:C.R.C.-July 1964

OTHER SPECIES PRESENT: helloides, editha, and rubidus.
MISC. DATA: Benchmark at base of ravine is BM4950'. Dock was common in marshy spots along the creek and along the road. specimens were visiting goldenrod.
CODED RESEARCH DATA REQUESTED BY SPECIALISTS.

TAXO. X in col. 45	29 2 3 5 6 8 4 4 5 2 1 6 4 2 3 1 1 0 5	46
ECOL. 12 in col. 72	72 2 1 4 4 5	80
CLIM. X in col. 47	47 2 3 1 1 0 3 2 6 5 1 1 3 7 2 5 0	63

Explanation of Figure

Fig. 3. The suggested field data card format (not copyrighted). With this type of card, all raw data concerning butterflies will be collected and punched on a standard punch card format.

On some of the more controversial genera, as in *Speyeria*, *Euphydryas*, and *Colias*, areas of common ground will have to be agreed on, so that a number may be applied to any given specimen. There, of course, will be some specimens that resist numbering, but probably can be broken down statistically and defined by numerical taxonomic methods and procedures.

Species

Some systematists would attempt to consolidate various species and form names (Hovanitz, 1943), ignoring variants from the original description or types. Others, more versatile, even attack basic species concepts (Ehrlich, 1961a). Realizing that probably both have some useful application, it is necessary for automation to approach the recording area with a general compromise.

As the primary usefulness of data processing lies in standardization of all values to be coded, it is noticed that members of many genera are quite resistant to any form of pigeonholing. It would appear from some recent literature (Ehrlich, 1961a, b; Ehrlich & Holm, 1962) that a species or subspecies is only a signpost along a long trail, and the post is planted only at the convenience of the original describer of the first specimen. Ehrlich and Holm believe this type of thinking is now slowly on the decline as it becomes more obvious that the clear-cut "biological species" concept is nonexistent in many cases, with the distinctness of many of our own butterfly species probably being vastly overrated.

Recently Ehrlich demonstrated a method of discriminating specimens by comparing their taxonomic features. That is, *Euphydryas editha colonia* would no longer be a name, but a concept expressed with a specific number representing coded values of certain taxonomic characters. These numbers, in themselves, would replace conventional methods of naming of various supposedly "distinct" entities found in the field. Names, then, will be retained only as a convenience in speaking or writing of a very wide group.

Ehrlich (1961a) after a careful analysis of 74 different characters on 13 specimens of *Euphydryas*, was able to construct with automated equipment (Burroughs 220, checked on IBM 650 series) a statistical diagram that revealed significant discriminatory information of the specimens, and established a base for speculaton on origins. It is probable that in most cases, after a preliminary investigation of great numbers of characters with each taxon, the total number might be simplified and standardized to those characters that are most significant. Stroud pioneered in this statistical area using only 14 characters in a study of termites (Sokal & Michener, 1958). More, perhaps, would be necessary to accumulate significant data with most of our genera, but the possibilities that novice collectors, with simplified procedures and equipment, might be of value in this field are great.

The present literature offers many examples of standardization suitable for widespread usage with automated equipment. Hovanitz (1943) has formulated a table for California *Speyeria callippe*, which need only be extended to other *callippe* forms, and addition of other characteristics than wing color patterns. In conjunction with Jude Le Gare (1951), he diagrammed and coded the pattern elements of *Melitaea chalcedona*, which may be applied to any form of *Euphydryas*, again adding other characteristics to better represent the known variation. These are only two of the many noted examples. By compiling, editing, and coding variable features of a species, then, eventually standardized procedures may be developed with which one may adequately deal with any butterfly variation found in the field or produced in the laboratory.

A notable quote from Ehrlich (1961a): "The continued presence of authors' names following the names of species of North American butterflies is, in most cases, a waste of type." "Citation of authors' names as a matter of course should cease." To my way of thinking, the present dos Passos checklist (1964) should be used as a standard directory, ignoring authors, until such time as a truly knowledgable and meaningful change in the status of the names may be made. With governing by the International Commission on Zoological Nomenclature, the rules and names comprise a usable system, but perhaps another, more usable system, should replace it.

Some Advantages of Automation

A great number of existing possibilities for every collector and interest is apparent, besides having all records available even though files or collections may be lost through accident or demise.

New specialists would be attracted to the field, making contributions from their own knowledge and specializations. Mathematicians and statisticians will be involved directly in the project, adding their training and experience to that of the taxonomist.

Of necessity, the record-keeping practices of participating collectors will improve. Vast amounts of information normally ignored, or lost in field notebooks, will come to light.

With a listing of all localities that have been collected, "blank" areas of previously uncollected areas will be noted, and perhaps collected, adding new distributional data to the files.

The only requirement of card punching, as mentioned earlier, is standardization of data, so that any conceivable type of data may be preserved on the punched card and compared in a statistical analysis with any other type of data.

With latitude and longitude accurately determined, it is possible to punch other sets of cards with extensive ecological data: climatological means, soil data, solar radiation maps, radioactive background counts, foliage cover and plant or tree distributions, geological maps, and other type of information found in a standardized format that could be carded. This data could be collated with that of the butterfly specimens to yield considerable information on life habits, habitats, and distributions, besides giving a means of comparing taxonomic features with ecological data to find, possibly, a previously unknown or unsuspected correlaton.

The long, tedious, computations necessary in correlating could be carried out rather simply with automated equipment. Large volumes of data could be automatically correlated as a matter of course, resolving obscure questions on pattern and ecology relationships, or other similar problems.

Presently, the U. S. Navy has ventured into electronography (Plain, 1964), with electronic taping of entire pieces of literature. With their present equipment, some 17,400 characters are composed per second, with enlarged microfilm negatives being printed at the rate of 240 an hour. This, in the future, then, would allow a specialist to microfilm every collected specimen and present it to the computer. At the press of a button, every known fact about that species could be at hand, along with photographs of every specimen recorded. Fantastic, I admit, but such a process is well within the realm of possibility.

Some Disadvantages of Automation

Fred Thorne (1964) writes that in the past, development of the Annual Season Summary series has had many difficulties, both in getting members to cooperate and in getting them to use the now standardized format. How, then, would members cooperate in the gigantic task of accumulating the necessary data for automation? A difficult question, which could be answered during the eventual trial of automated data processing.

The IBM equipment cannot perform miracles, and even the simplest operation may consume considerable time. This would necessitate research priorities, and possibly long waiting lists. To circumvent general requests of a distributional nature, the task of preparing the Annual Season Summary could be assumed by the machines, only reporting on a much wider scope, giving species listings, or maps, of every specimen reported during that year. After sufficient data is collected and compared with ecological data, distribution maps of probable range could be easily issued. These maps could then be accumulated by collectors, and used to solve most distributional problems.

Needless to say, many "specimen hunters" or commercial collectors will take neither the time nor effort to support a project of automated nature.

Summary

It is inevitable that one day a system similar to that described shall have to be initiated in this field. Presently, collections are becoming so vast, and scattered among so many institutions and individuals, that any major study is becoming quite difficult, through the sheer volume of data to be accumulated and processed.

If a forecast were required, I should call for a vast system of interlocked computers, handling the new information as it is collected and correlating the mass with giant stores of previous entries. We can probably look forward to an institution that determines individual collectors' annual catch, for the privilege of microfilming the specimens and submitting the data to the computer. The vast amount of literature concerning butterflies will be scanned and all relevant information and facts taped or stored in the computers—the factual data being emitted in a single lump at the push of a button. The system will spread to encompass not only butterflies, but all insects, plants, and animals, present and extinct.

Quite probably, before the end of the next few years stimulating entomological papers of significance will be forthcoming from those not directly interested in entomology, but in statistics, mathematics, and automated equipment. While emphasis on nomenclature sinks to obscurity, data emitted from the bowels of an electronic maze of transistors, wires, and memory cells will formulate new concepts concerning behavior, comparative anatomy, genetics, geographical and ecological distribution, and evolutionary trends.

The author was graduated from the Western Automation Institute in Portland, Oregon, and for a year and a half before entering the meteorology field, operated the equipment mentioned in this report.

I wish to offer my thanks to L. P. Grey, Dr. J. A. Powell, and F. T. Thorne for their invaluable aid with criticisms and suggestions on this paper.

Literature Cited

Anonymous, 1960. General Information Manual, Introduction to IBM Data Processing Systems. I.B.M. Corp.

1961. General Information Manual, IBM 1401 Data Processing System from Control Panel to Stored Program. I.B.M. Corp.

Arkin, H., & R. R. Colton, 1964. Statistical Methods. Barnes & Noble, Inc., New York.

Brown, C. E., 1964. A machine method for mapping insect survey records. Contr. 1103, Forest Ent. & Pathol. Branch, Dept. Forestry, Ottawa, Canada.

Brown, W. L., Jr., 1964. Pilot Register of Zoology. N. Y. College of Agriculture at Cornell Univ.

Copeland, J. A., 1949. The Weather Bureau Punched Card Project. 1946–1947 Lectures to Professional Interns. Weather Bureau Training Paper No. 2: 35–48.

dos Passos, C. F., 1964. A synonymic list of the Nearctic Rhopalocera. Lep. Soc. Memoir, No. 1.

Ehrlich, P. R., 1961a. Has the biological species concept outlived its usefulness? Systematic Zoology, 10(4): 167–176.

1961b. Systematics in 1970: Some unpopular predictions. Systematic Zoology, 10(4): 157–158.

Ehrlich, P. R., & R. W. Holm, 1962. Patterns and populations. Science, 137(3531): 652–657.

Garrison, G. A., & J. M. Skovlin, 1960. Northwest range plant symbols. Pac. N.W. Forest & Range Exp. Sta., U. S. Dept. of Agriculture–Forest Service.

Hovanitz, W., 1943. Geographical variation and racial structures of *Argynnis callippe* in California. Amer. Nat., 77: 400–425.

JUDE LE GARE, M., & W. HOVANITZ, 1951. Genetic and ecologic analyses of wild populations in Lepidoptera. II. Color pattern variation in *Melitaea chalcedona.* Wasmann Jour. Biol., 9(3): 257–310.

McGUGAN, B. M., 1958. Forest Lepidoptera of Canada. Vol. 1. Papilionidae to Arctiidae. Canad. Dept. Agric. Publ. 1034: 1–76.

PLAIN, C., 1964. Writers Envision Automated Library. San Diego Union. May (day of issue unknown).

POWELL, J. A., 1965. Personal communication.

PRENTICE, R. M., 1962. Forest Lepidoptera of Canada. Vol. 2. Nycteolidae to Liparidae. Canad. Dept. Forestry, Bull. 128: 77–282.

1963. Forest Lepidoptera of Canada. Vol. 3. Lasiocampidae to Geometridae. Canad. Dept. Forestry, Publ. 1013: 283–543.

ROHLF, F. J., 1963a. Classification of *Aedes* by numerical taxonomic methods. Ann. Ent. Soc. Amer., 56(6): 798–804.

1963b. Congruence of larval and adult classifications in *Aedes*. Systematic Zoology, 12(3): 97–117.

SOKAL, R. R., & C. D. MICHENER, 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38, Part II(22): 1409–1438.

SOKAL, R. R., & P. H. A. SNEATH, 1963. Principles of Numerical Taxonomy. W. H. Freeman & Co., San Francisco.

ULAM, S. M., 1964. Computers. Sci. Amer., 211(3): 202–216.

URNESS, D. B. (IBM Systems Engineer), 1964. Personal communication.

CERCYONIS PEGALA NEPHELE (SATYRIDAE) AT FLUORESCENT LIGHT

Having read the recent notices of Rhopalocera taken at light, both in the *Journal of the Lepidopterists' Society* and the *Entomologist's Record and Journal of Variation*, I was most interested to find another species attracted to fluorescent light. While collecting Noctuidae in the company of John Newman at Morenci, Michigan, on the evening of July 31, 1965, using a 15-watt "BL" fluorescent black light suspended before a white sheet on a frame, a male *Cercyonis pegala nephele* (Kirby) was seen to fly at the lighted sheet and react in the same manner as a nocturnal insect. It was obviously attracted by the light, and soon settled upon the ground flap.

Due to the location of the apparatus and other circumstances, it is safe to say that the butterfly was not mechanically disturbed from its resting place, but was actually drawn by the lamp. We had seen the species occasionally while collecting in the afternoon. The specimen was taken at approximately 10:30 P.M., and a light rain of over two hours' duration had just ceased, reinforcing the certainty of attraction.

RONALD S. WILKINSON, 2337 *Hamilton Rd., Okemos, Michigan*