Dallwitz, M.J. 1974. A flexible computer program for generating identification keys. Syst. Zool. 23:50–57. Also available at www.delta-intkey.com/www/dallwitz-1974.pdf.

A flexible computer program for generating identification keys

1974

M.J. Dallwitz

Abstract

A Fortran program for constructing identification keys is described. The strategy used by the program aims to minimize the probability of error or the amount of difficulty in using the key. The main considerations in assessing the merit of a character are the reliability or convenience of the character (as estimated by the maker of the key), the evenness of the division produced by the character, and the amount of intra-taxon variability for the character. The key maker has a high degree of control over the strategy, as the relative importance of these criteria may be specified, and the automatic choice of characters may be overridden.

Introduction

The construction of identification keys by computer has many advantages over construction by hand, especially for large groups of taxa. The labor involved in collecting the data is still considerable, but once this has been done, the production of keys is so simple that it is quite practical to make special keys for laboratory or field identification, or for particular localities or parts of the life cycle, and keys can easily be revised if errors are found or new taxa or characters have to be added. Furthermore, keys produced by computer are usually better than those produced by hand, because the characters are chosen systematically according to some rule that aims to optimize the structure of the key.

This paper describes a new computer program, called KEY, for generating identification keys. The program has several advantages over those described by Hall (1970), Pankhurst (1970), and Morse (1971).

The section describing the character-selection algorithm is not essential to the understanding of the rest of the paper.

Terminology

The individuals to which a key is to be applied will be grouped into two types of class: (1) Taxon—a class of individuals with the same name (at the taxonomic level for which the key is required). The purpose of the key is to assign any given individual to its taxon. (2) Item—a class of individuals with identically coded descriptions. A taxon may include more than one item because of variation within the taxon.

The description of a group of taxa will be segmented into descriptions of homologous features or properties. A feature, together with an indication of the type of description to be applied to it, will be called a character (e.g., color of wings). The different ways in which a character may be expressed in the group will be called the states of the character, and each state will be identified by a letter or number (e.g., A. white, B. yellow, C. brown). The identification letter or number of the character state describing a particular item will be called the value of the character for the item. This term will also be used to denote symbols indicating that a character is not applicable to an item or is variable for an item. A character together with a value will be called an attribute.

General description of the program

The program user provides a list of characters and character states, and a table giving the taxa corresponding to the items and the values of the characters for the items. Missing values may be interpreted as not applicable or as variable. The program user may also provide estimates of the costs (in the wide sense) of using the characters and of the relative abundances of the items. (By default, all costs and all abundances are assumed to be equal.) An interpretation of the meaning of the cost of a character is not essential for the use of the program or for the development of the theory, but it may be thought of as a combination of the probability of error in using the character and the amount of effort required to use the character. (Errors may arise because of misinterpretation or misjudgment by the user of the key, or because intra-taxon variability has not been completely accounted for in the taxon descriptions.)

Under the assumptions that the character costs are additive (that is, that the cost of using two characters is the sum of the costs of the characters), and that the frequency with which an item will need to be identified is proportional to its abundance, the program attempts to construct a key in which the average cost of an identification is minimized. The key is printed in a tabular form resembling a tree diagram, or in the conventional bracketed style.

The construction of a key by the program proceeds in the following way. The ‘best’ character is selected, and is used to divide the original group of taxa into two or more subgroups. Each subgroup is then subdivided by the best remaining character (not necessarily the same one for each subgroup), and this process is continued until each subgroup contains only one taxon (or until no suitable character can be found, in which case the key will be incomplete).

Instead of allowing the program to choose all the characters for the key, the key maker may specify (preset) the characters to be used at any positions in the key. There are two possible reasons why this may be desirable. Firstly, the key maker may want the grouping of taxa in the key to reflect taxonomic relations, and this may not be achieved if the characters are selected by the program. Secondly, the strategy used by the program does not necessarily produce the best possible key, and it may be possible to improve the key by presetting characters.

Provision is made for masking of both characters and taxa; that is, keys may be formed from subsets of the data without altering the data cards. This makes possible the easy construction of special-purpose keys—for example, a key for a particular locality, or a field key using characters not requiring dissection or microscopic examination.

The character-selection algorithm

The first step in the choice of the best character for the subdivision of a group of taxa is to eliminate those characters which take only one value in the group, or are not applicable to some members of the group. A ‘comparison function’ is then evaluated for each remaining character, and the character giving the lowest value is chosen.

The comparison functions used by Hall (1970), Pankhurst (1970), and Morse (1971), were chosen to favor characters of low cost and characters which most nearly divide the group into equal subgroups. The first of these criteria is obviously desirable, and the second is based on some results of Osborne (1962). However, all of these authors use different functions, and do not attempt to justify the detailed form of the function. The comparison function used in KEY is an estimate of the average cost of an identification in the subkey under consideration. It was derived by the following argument.

We define the length L of a key as the average number of questions that must be answered to identify a taxon; that is,

$$L = (1/n)\sum_{i=1}^{n} l_i$$

where n is the number of taxa and l_i is the number of questions required to identify the i-th taxon. It is easily verified that if all the characters have two states, the minimum value of L over all conceivable keys is given by

$$L_{\min} = k + 2 - (2^{k+1}/n) \quad (1)$$

where k is the integral part of $\log_2 n$. The value of $L_{\min}$ may be approximated by the simpler formula

$$L_{\min} \approx \log_2 n$$

which is exact if n is a power of 2, and differs from the value given by equation (1) by at most 0.086.

Suppose a character having a cost c is used to divide a group of n taxa into s subgroups. A key for the group can then be obtained by constructing subkeys for each of the subgroups. The average cost C of an identification in the key is given by

$$C = c + \left(\sum_{j=1}^{s} a_j c_j L_j\right) / \left(\sum_{j=1}^{s} a_j\right) \quad (2)$$

where c_j is the 'average' cost of the characters used in the j-th subkey, a_j is the total abundance of the taxa in the j-th subgroup, and L_j is the length of the j-th subkey. If we further assume that the subkeys are not far from minimum length, and that most of the characters have two states, equation (2) becomes

$$C \approx c + \left(\sum_{j=1}^{s} a_j c_j \log_2 n_j\right) / \left(\sum_{j=1}^{s} a_j\right) \quad (3)$$

where n_j is the number of taxa in the j-th subgroup.

The estimation of the c_j before the sub-keys are constructed is very difficult, and in view of the many other approximations, an elaborate procedure is not warranted. In the program, all the c_j are simply set equal to $c_{\min}$, the smallest cost for the characters under consideration. Equation (3) then becomes

$$C \approx c + c_{\min}\left(\sum_{j=1}^{s} a_j \log_2 n_j\right) / \left(\sum_{j=1}^{s} a_j\right) \quad (4)$$

The comparison function K used in the program is obtained from equation (4) by adding a term which allows the user to control the amount of intra-taxon variability in the key:

$$K = c + c_{\min}\left[\left(\sum_{j=1}^{s} a_j \log_2 n_j\right) / \left(\sum_{j=1}^{s} a_j\right) + ((1 - v)/v)(1 + 100m)\left(\sum_{j=1}^{s} n_j - n\right) / \log_2 n\right] \quad (5)$$

where v is the absolute value of the parameter VARYWT described in the next section and m is the number of subgroups for which $n_j = n$. It should be noted that the extra term in equation (5) is an arbitrary one, unrelated to the minimization of the cost. Even without this term, the comparison function favors characters with low intra-taxon variability (unlike the function used by Pankhurst (1970)).

It is interesting that for the case when all the costs are equal, all the abundances are equal, $s = 2$, and there is no intra-taxon variability, a comparison function based on equation (4) reduces to

$$K = n_1 \log_2 n_1 + n_2 \log_2 n_2$$

a result obtained by Maccacaro (1958) and Shwayder (1971) from information theory.

The comparison function in equation (5) tends to favor multistate characters over two-state characters, and yet most taxonomists prefer two-state characters in keys. There are probably two main reasons for this preference. Firstly, in the indented style of key, the descriptions of the states may be widely separated, and the key is easier to use if it is known that each character has only two states. If this style of key is required, the best approach is to use only two-state characters in the original character list. Secondly, it is felt that errors are more likely in determining the values of multistate characters, especially if the characters in question involve long or complicated descriptions. Where this is the case, the character can be given a higher cost, or can be split up into several two-state characters. However, in many cases (see Fig. 1), multistate characters are no more difficult to use (sometimes less so) than the equivalent two-state characters, and their use can lead to considerably shorter keys.

Using the program

Examples of output from the program are shown in Figs 1 to 5. The data are part of a larger set comprising descriptions of 135 species in terms of 40 characters. In these examples, the character states and values are identified by letters, but numbers can be used if preferred.

```
1. (8) NUMBER OF ANTENNAL SEGMENTS
          A. NINE
          B. EIGHT
2. (8) NUMBER OF LAMELLAE IN ANTENNAL CLUB
          A. THREE
          B. FOUR
          C. FIVE
          D. SIX
3. (6) HEAD COLOUR
          A. BLACK
          B. REDDISH
          C. YELLOWISH BROWN
4. (8) LATERAL MARGINS OF PRONOTUM
          A. GRANULAR AND WITH NUMEROUS SETAE
          B. NOT GRANULAR AND WITH A SINGLE ROW OF SETAE
5. (6) LATERAL EDGES OF PRONOTUM
          A. CONVEX BEFORE POSTERIOR ANGLES
          B. STRAIGHT OR CONCAVE BEFORE POSTERIOR ANGLES
6. (5) PRONOTUM SURFACE
          A. SHINING
          B. IRIDESCENT
          C. DULL
7. (4) PRONOTUM COLOUR
          A. BLACK
          B. REDDISH
          C. YELLOWISH BROWN
          D. BICOLOURED
8. (7) PROPYGIDIUM
          A. WITH A DENSE FRINGE OF SETAE CLOSE TO THE POSTERIOR EDGE
          B. WITHOUT A FRINGE OF SETAE
9. (7) LENGTH IN MM
          A. LESS THAN 10
          B. BETWEEN 10 AND 15
          C. BETWEEN 15 AND 20
          D. BETWEEN 20 AND 25
          E. BETWEEN 25 AND 30
          F. GREATER THAN 30
10. (5) RATIO OF LENGTH TO WIDTH OF MAXILLARY PALP
          A. LESS THAN 2
          B. BETWEEN 2 AND 4
          C. BETWEEN 4 AND 6
          D. GREATER THAN 6
```

Fig. 1. List of characters, states, and reliability indices for species of *Colpochila* (Coleoptera: Scarabaeidae: Melolonthinae).

Fig. 1 is a list of the characters and states used in the descriptions. The numbers in brackets are ‘reliability indices’ that indicate the preferences of the key maker for the characters. The reliability indices may be measures of the reliability, convenience, or accessibility of the characters, or may indicate preferences for particular classes of characters, such as vegetative or larval characters.

```
 1. (5) BICOLOR BLKB.        AAABB BABBB
 2. (5) ELECTA BLKB.         AABBB ABACB
 3. (5) ERYTHROCEPHALA LEA   BABBA CDADC
 4. (5) FIMBRIATA LEA        BABBA CBACB
 5. (5) FIMBRICOLLIS LEA     AABAA ABADA
 6. (4) GAGATINA BURM.       AAABA CABBB
 7. (4) GAGATINA BURM.      *         C
 8. (5) GYMNOPYGA LEA        AABBV BBBDB
 9. (5) NEW SPECIES A        ABCBA ACACB
10. (5) NEW SPECIES B        ABBBA BBBEB
```

Fig. 2. List of taxa, character values, and abundance indices for species of *Colpochila*.

Fig. 2 is a list of taxa and character values. The description of NEW SPECIES A, for example, would read as follows: nine antennal segments, four lamellae in antennal club, yellowish brown head, etc. The numbers in brackets are ‘abundance indices’ which are assigned to the items by the maker of the key. Items with higher abundance indices will tend to come out earlier in the key than those with lower abundance indices. An abundance index would normally be a measure of the abundance of the individuals of the item, so that the more common items could be identified more quickly.

The character values ‘–’ and ‘V’ have special meanings. A value of ‘–’ means that the character is not applicable to the item, and will not be used for divisions involving that item. A value of ‘V’ means that the character is variable or poorly defined for the item. In this case, it is assumed that the character may take any of its possible values (except ‘–’), and if the character is used for the item, the corresponding taxon will appear more than once in the key. Unknown values are left blank and are treated as not applicable or variable, as specified by the program user.

Intra-taxon variability may also be expressed by inserting more than one item for the taxon. This is advisable when the taxon does not exhibit the full possible range of variability of a character, or when two or more variable characters are correlated. In the second and subsequent items of a taxon, only those values that differ from the first item need be filled in (Watson and Milne 1972). (A flag is required on the input card to indicate that the blanks do not represent unknown values.) For example, the two items for GAGATINA in Fig. 2 indicate that the length is between 10 mm and 20 mm (attribute 9B or 9C).

```
+---------------------------+---+---+---+
|FIMBRICOLLIS LEA           | 2A| 1A| 4A|
+---------------------------+   |   |---+---+
|ELECTA BLKB.               | 2A| 1A| 4B| 8A|
+---------------------------+   |   |   |---+---+---+
|GAGATINA BURM.             | 2A| 1A| 4B| 8B| 3A| 5A|
+---------------------------+   |   |   |   |   |---+
|BICOLOR BLKB.              | 2A| 1A| 4B| 8B| 3A| 5B|
+---------------------------+   |   |   |   |---+---+
|GYMNOPYGA LEA              | 2A| 1A| 4B| 8B| 3B|
+---------------------------+   |---+---+---+---+
|FIMBRIATA LEA              | 2A| 1B| 9C|
+---------------------------+   |   |---+
|ERYTHROCEPHALA LEA         | 2A| 1B| 9D|
+---------------------------+---+---+---+
|NEW SPECIES A              | 2B| 9C|
+---------------------------+   |---+
|NEW SPECIES B              | 2B| 9E|
+---------------------------+---+---+
```

Fig. 3. A key to species of *Colpochila*, with high weight on character reliability and on minimization of intra-taxon variability.
Parameter values: RBASE = 2, ABASE = 2, REUSE = 1.01, VARYWT = 0.1, NCONF = 0.

A key produced from these data is shown in Fig. 3. To identify a specimen, its characters are examined in the order specified in the key, reading from left to right. As the value of each character is determined, the group of possible taxa is reduced, until finally only one taxon remains, and the identification is complete. For example, an identification using the key in Fig. 3 might proceed as follows. The first character to be examined is 2—number of lamellae in antennal club. Suppose the specimen has three lamellae, that is, attribute 2A. Then the specimen cannot be NEW SPECIES A or NEW SPECIES B, and must be one of the other seven species. The next character is 1— number of antennal segments. If the specimen has eight antennal segments (attribute 1B), it is either FIMBRIATA or ERYTHROCEPHALA. Finally, we examine character 9— length. If the specimen is 16 mm long (attribute 9C), it is FIMBRIATA.

```
+---------------------------+---+---+---+
|GAGATINA BURM.            :| 9B| 5A  6C|
+---------------------------+   |---+---+
|BICOLOR BLKB.              | 9B| 5B  6B|
+---------------------------+---+---+---+
|GAGATINA BURM.            :| 9C| 3A  7A|
+---------------------------+   |---+---+---+---+
|ELECTA BLKB.               | 9C| 3B  7B| 1A  5B|
+---------------------------+   |       |---+---+
|FIMBRIATA LEA              | 9C| 3B  7B| 1B  5A|
+---------------------------+   |---+---+---+---+
|NEW SPECIES A              | 9C| 3C  7C|
+---------------------------+---+---+---+
|FIMBRICOLLIS LEA           | 9D| 6A 10A|
+---------------------------+   |---+---+
|GYMNOPYGA LEA              | 9D| 6B 10B|
+---------------------------+   |---+---+
|ERYTHROCEPHALA LEA         | 9D| 6C 10C|
+---------------------------+---+---+---+
|NEW SPECIES B              | 9E|
+---------------------------+---+
```

Fig. 4. A key to species of *Colpochila*, with confirmatory characters and low weight on character reliability and on minimization of intra-taxon variability.
Parameter values: RBASE = 1.1, ABASE = 2, REUSE = 1.01, VARYWT = 0.9, NCONF = 1.

Fig. 4 is another key, produced from the same set of data by changing the values of certain parameters. This key illustrates the use of confirmatory or alternative characters. Where adjacent columns are not separated by vertical ruling, the characters in these columns are equivalent in that their values have the same distributions in the group of taxa under consideration, and either or both may be used in an identification. Such characters are arranged in order of decreasing reliability (from left to right).

The strategy used to select characters for the key is controlled by the parameters RBASE, ABASE, REUSE, VARYWT, and NCONF. By varying these parameters, the program user can produce different keys from the same set of data.

The parameter RBASE controls the weight given to the reliability indices. From the reliability index of each character, the program calculates the cost of the character as

$$\text{cost} = \text{RBASE}^{-(\text{reliability index})}$$

Thus, if RBASE = 1, all characters have equal costs, and the reliability indices have no influence on the formation of the key. If RBASE = 2, each increase of 1 in the reliability index means a halving of the cost. The method by which the program chooses characters is such that when RBASE = 2, a more reliable character will almost always be used in preference to a less reliable character, as in Fig. 3. This is not necessarily desirable, as it tends to increase the number of questions which must be answered in making an identification. The shortest keys are obtained when RBASE is not much greater than 1, as in Fig. 4, but this may be at the expense of using less reliable characters.

ABASE controls the effect of the abundance indices on the structure of the key. The relative abundance of an item is given by

$$\text{abundance} = \text{ABASE}^{(\text{abundance index})}$$

The program takes the abundances into account in assessing the evenness of the subdivision produced by a character. Abundant items tend to be placed in groups which are small in terms of the number of taxa they contain, with the result that abundant items tend to come out early in the key. The emphasis given to this effect is determined by the value of ABASE.

When a character is first used in the key, its cost is divided by REUSE. If REUSE is greater than 1, the character is then more likely to be used again later in the key. The aim of this is to minimize the number of characters used in the key. As this is not usually an important requirement, the value of REUSE should usually be only slightly greater than 1.

The value of VARYWT determines the treatment of characters which are variable for some taxa, and the treatment of unknown values. If VARYWT = 0, characters with intra-taxon variability will not be used for decisions involving the variable taxa, and each taxon name will appear only once in the key. If |VARYWT| = 1, there is no special penalty for such characters (although there is still a tendency to avoid them). Intermediate values of |VARYWT| have effects between these extremes. Unknown values are treated as not applicable if VARYWT is positive, and variable if VARYWT is negative. The keys in Figs. 3 and 4 illustrate the effect of VARYWT. In Fig. 3, a low value of VARYWT results in the avoidance of character 9 for groups containing the species GAGATINA, which is variable for this character. In Fig. 4, VARYWT is large, and character 9 is the first one used, with the result that GAGATINA appears twice in the key. The tendency of the program to avoid characters with intra-taxon variability is especially valuable when this variability is high, as is often the case when dealing with taxonomic categories higher than the species. For example, in the data of Watson and Milne (1972), 66 grass genera of the Australian Capital Territory are represented by 245 items; keys produced for these genera contained 80 items when VARYWT was 0.9 and 86 items when VARYWT was 1.

NCONF is the maximum number of confirmatory characters to be sought for each main character.

```
1(0). LENGTH IN MM BETWEEN 10 AND 15................................ 2
      LENGTH IN MM BETWEEN 15 AND 20................................ 3
      LENGTH IN MM BETWEEN 20 AND 25................................ 5
      LENGTH IN MM BETWEEN 25 AND 30.................... NEW SPECIES B

2(1). LATERAL EDGES OF PRONOTUM CONVEX BEFORE POSTERIOR ANGLES;
        PRONOTUM SURFACE DULL......................... GAGATINA BURM.
      LATERAL EDGES OF PRONOTUM STRAIGHT OR CONCAVE BEFORE POSTERIOR
        ANGLES; PRONOTUM SURFACE IRIDESCENT............. BICOLOR BLKB.

3(1). HEAD COLOUR BLACK; PRONOTUM COLOUR BLACK......... GAGATINA BURM.
      HEAD COLOUR REDDISH; PRONOTUM COLOUR REDDISH................. 4
      HEAD COLOUR YELLOWISH BROWN; PRONOTUM COLOUR YELLOWISH BROWN....
         ............................................. NEW SPECIES A

4(3). NUMBER OF ANTENNAL SEGMENTS NINE; LATERAL EDGES OF PRONOTUM
        STRAIGHT OR CONCAVE BEFORE POSTERIOR ANGLES...... ELECTA BLKB.
      NUMBER OF ANTENNAL SEGMENTS EIGHT; LATERAL EDGES OF PRONOTUM
        CONVEX BEFORE POSTERIOR ANGLES.................. FIMBRIATA LEA

5(1). PRONOTUM SURFACE SHINING; RATIO OF LENGTH TO WIDTH OF MAXILLARY
        PALP LESS THAN 2............................ FIMBRICOLLIS LEA
      PRONOTUM SURFACE IRIDESCENT; RATIO OF LENGTH TO WIDTH OF
        MAXILLARY PALP BETWEEN 2 AND 4.................. GYMNOPYGA LEA
      PRONOTUM SURFACE DULL; RATIO OF LENGTH TO WIDTH OF MAXILLARY
        PALP 4 BETWEEN AND 6...................... ERYTHROCEPHALA LEA
```

Fig. 5. The key in Fig. 4, printed in the conventional bracketed style.

The type of key shown in Figs 3 and 4 is convenient for examining the overall structure of the key, for checking the routes by which individual taxa are brought out, and for comparison of different keys produced from the same data. The program can also produce a key with a renumbered character list containing only those characters actually used in the key, or a conventional bracketed style of key (see Fig. 5).

A major advantage of KEY is its low cost. For example, a tabular key to 145 Australian grass genera was produced from 433 items and 120 characters in 80 seconds of computer time (on a CDC3600) at a cost of $5 (cf. Watson and Milne 1972, Pankhurst 1970). Printing a conventional key took 25 seconds more and cost $2 more.

Details of the program

Two versions of the program are available. One is written mainly in ANSI Fortran (American National Standards Institute, 1966), and the other is written in CDC3600 Fortran and Compass. The non-ANSI part of the first version comprises two simple character-handling subroutines (to store and load the n-th character of an array), and is not needed unless keys printed in the bracketed style are required. The first version has run successfully on a UNIV AC 1108.

The coded descriptions of the items are stored as a matrix, with rows corresponding to items and columns to characters. Each element of the matrix occupies 12 bits: 8 bits contain the character number, 3 bits contain the character value, and the remaining bit is a flag. (Initially, character numbers are the same as column numbers, but this relationship is not preserved.) Fortran subroutines are supplied to store and retrieve the data, with one element per computer word. These may be replaced by assembly language subroutines storing several elements per word.

The character and state descriptions are stored only if a key with renumbered characters or a bracketed key is required. Fortran subroutines are supplied to store these descriptions in core, but they may be replaced by subroutines storing the descriptions on random-access disc or drum, to reduce core requirements.

The key structure is stored within the data matrix itself by rearranging the elements. The column segments (characters) to be used in the key are collected in the left-hand side of the matrix, and the rows are sorted so that items with the same attributes are grouped together. If the value ‘V’ (variable) comes into the key, the program generates new rows (items) for all the possible variants, and inserts them in the matrix. The printed key is a direct printout of the relevant part of the matrix, with labeling and ruling added.

The Fortran program contains about 1300 statements. When this is compiled on a CDC3600, the program occupies about 9200 48-bit words, apart from blank common, which is used for data storage, and may be expanded to use all the remaining core (22800 words on the CSIRO CDC3600). A problem with c characters masked in and i items masked in requires roughly

$$ci/p + 18i + 7c$$

words for data storage, where p is the number of data elements packed in each word. In the CDC3600 version of the program, p is 4.

A writeup and listing of KEY are available from the author. Also, a program is available for converting the free-field data format developed by Watson and Milne (1972) to the format required by KEY.

Conclusion

The key-forming program which has been described has the following advantages. The strategy used by the program to select characters has been justified theoretically; it is flexible, allowing the key maker to specify the relative importance of character reliability, quick identification of common taxa, minimization of the number of characters used, and minimization of intra-taxon variation. Furthermore, the key maker may override the automatic selection of characters by specifying the characters to be used at any positions in the key, and may choose to have uncoded attributes treated as ‘not applicable’ or as ‘variable.’ Keys can be printed in the conventional bracketed style, or in a tabular form which displays the structure of the key clearly. The method of coding intra-taxon variation is compact yet flexible. The program is cheap to run, and there is provision for the use of backing store and partial words to handle large problems.

Acknowledgments

I thank E.B. Britton (Division of Entomology, CSIRO) for providing test data, and for discussion and suggestions during the development of the program; L. Watson (Research School of Biological Sciences, Australian National University) and P. Milne (Division of Computing Research, CSIRO) for a preprint of their paper, discussion, and access to data; and D.H. Colless and K.H.L. Key (Division of Entomology, CSIRO) for critical reading of the manuscript.

References

American National Standards Institute. 1966. American National Standard FORTRAN, X3.9–1966. American National Standards Institute, New York.

Hall, A.V. 1970. A computer-based system for forming identification keys. Taxon 19, 12–18.

Maccacaro, G. A. 1958. La misura della informazione contenuta nei criteri di classificazione. Ann. Microb. Enzimol. 8, 231–239.

Morse, L.E. 1971. Specimen identification and key construction with time-sharing computers. Taxon 20:269-282.

Osborne, D. V. 1963. Some aspects of the theory of dichotomous keys. New Phytol. 62, 144–160.

Pankhurst, R. J. 1970. A computer program for generating diagnostic keys. Comp. Jour. 13, 145–151.

Shwayder, K. 1971. Conversion of limited-entry decision tables to computer programs—a proposed modification of Pollack's algorithm. Comm. ACM 14, 69–73.

Watson, L., and Milne, P. 1972. A flexible system for automatic generation of special-purpose dichotomous keys, and its application to Australian grass genera. Aust. J. Bot. 20, 331–352.

Manuscript received January, 1973. Revised September, 1973

Appendix

10 March 2020

This appendix is not part of the original paper. It shows versions of the data and keys as used or generated by the current version of Key and the related program Confor. The data are now recorded in DELTA format (Dallwitz 1980; Dallwitz, Paine and Zurcher 1993 onwards), and are converted by Confor to the format used by Key.

```
*SHOW ~ Species of Colpochila – character list

*CHARACTER LIST

#1. antennae with <number of segments>/
      segments/

#2. antennal club with <number of lamellae>/
      lamellae/

#3. head <colour>/
      1. black/
      2. reddish/
      3. yellowish brown/

#4. pronotum lateral margins <granularity and setae>/
      1. granular and with numerous setae/
      2. not granular and with a single row of setae/

#5. pronotum lateral edges <whether convex>/
      1. convex before posterior angles/
      2. straight or concave before posterior angles/

#6. pronotum surface <reflectivity>/
      1. shining/
      2. iridescent/
      3. dull/

#7. pronotum <colour>/
      1. black/
      2. reddish/
      3. yellowish brown/
      4. bicoloured/

#8. propygidium <whether fringed with setae>/
      1. with a dense fringe of setae close to the posterior edge/
      2. without a fringe of setae/

#9. length <overall>/ mm/

#10. maxillary palp ratio of length to width/
```

Fig. 1a. Character list in DELTA format. Characters 1, 2, 9, and 10 are numeric. For use by Key, they are converted to multistate by Confor.

```
*SHOW ~ Species of Colpochila - item descriptions

*ITEM DESCRIPTIONS

# \i{}C. bicolor\i0{} <Blkb.>/
1,9 2,3 3,1 4,2 5,2 6,2 7,1 8,2 9,10.1-14.9 10,2.1-3.9

# \i{}C. electa\i0{} <Blkb.>/
1,9 2,3 3,2 4,2 5,2 6,1 7,2 8,1 9,15.1-19.9 10,2.1-3.9

# \i{}C. erythrocephala\i0{} <Lea>/
1,8 2,3 3,2 4,2 5,1 6,3 7,4 8,1 9,20.1-24.9 10,4.1-5.9

# \i{}C. fimbriata\i0{} <Lea>/
1,8 2,3 3,2 4,2 5,1 6,3 7,2 8,1 9,15.1-19.9 10,2.1-3.9

# \i{}C. fimbricollis\i0{} <Lea>/
1,9 2,3 3,2 4,1 5,1 6,1 7,2 8,1 9,20.1-24.9 10,0.1-1.9

# \i{}C. gagatina\i0{} <Burm.>/
1,9 2,3 3,1 4,2 5,1 6,3 7,1 8,2 9,10.1-19.9 10,2.1-3.9

# \i{}C. gymnopyga\i0{} <Lea>/
1,9 2,3 3,2 4,2 5,V 6,2 7,2 8,2 9,20.1-24.9 10,2.1-3.9

# New species A/
1,9 2,4 3,3 4,2 5,1 6,1 7,3 8,1 9,15.1-19.9 10,2.1-3.9

# New species B/
1,9 2,4 3,2 4,2 5,1 6,2 7,2 8,2 9,25.1-29.9 10,2.1-3.9
```

Fig 2a. Item descriptions in DELTA format.

```
RBASE = 2.00 ABASE = 2.00 REUSE = 1.01 VARYWT = 0.10
Number of confirmatory characters = 0

Average length of key = 3.8 Average cost of key = 0.8
Maximum length of key = 6 Maximum cost of key = 1.6

Character reliabilities 1-2,8 3,6 4,8 5,6 7,4 8,7 9,7.1

+----------------------------+---+---+---+
|C. fimbricollis             | 2A| 1A| 4A|
+----------------------------+   |   |---+---+
|C. electa                   | 2A| 1A| 4B| 8A|
+----------------------------+   |   |   |---+---+---+
|C. gagatina                 | 2A| 1A| 4B| 8B| 3A| 5A|
+----------------------------+   |   |   |   |   |---+
|C. bicolor                  | 2A| 1A| 4B| 8B| 3A| 5B|
+----------------------------+   |   |   |   |---+---+
|C. gymnopyga                | 2A| 1A| 4B| 8B| 3B|
+----------------------------+   |---+---+---+---+
|C. fimbriata                | 2A| 1B| 9C|
+----------------------------+   |   |---+
|C. erythrocephala           | 2A| 1B| 9D|
+----------------------------+---+---+---+
|New species A               | 2B| 9C|
+----------------------------+   |---+
|New species B               | 2B| 9E|
+----------------------------+---+---+
```

Fig. 3a. Tabular key with high weight on character reliability and on minimization of intra-taxon variability.

```
RBASE = 1.10 ABASE = 2.00 REUSE = 1.01 VARYWT = 0.90
Number of confirmatory characters = 1

Average length of key = 2.1 Average cost of key = 1.8
Maximum length of key = 3 Maximum cost of key = 2.5

Character reliabilities 1-2,8 3,6 4,8 5,6 7,4 8,7 9,7.1

+---------------------------+---+---+---+
|C. gagatina               2| 9B| 5A  6C|
+---------------------------+   |---+---+
|C. bicolor                 | 9B| 5B  6B|
+---------------------------+---+---+---+
|C. gagatina               2| 9C| 3A  7A|
+---------------------------+   |---+---+---+---+
|C. electa                  | 9C| 3B  7B| 1A  5B|
+---------------------------+   |       |---+---+
|C. fimbriata               | 9C| 3B  7B| 1B  5A|
+---------------------------+   |---+---+---+---+
|New species A              | 9C| 3C  7C|
+---------------------------+---+---+---+
|C. fimbricollis            | 9D| 6A 10A|
+---------------------------+   |---+---+
|C. gymnopyga               | 9D| 6B 10B|
+---------------------------+   |---+---+
|C. erythrocephala          | 9D| 6C 10C|
+---------------------------+---+---+---+
|New species B              | 9E|
+---------------------------+---+
```

Fig. 4a. Tabular key with confirmatory characters, and low weight on character reliability and on minimization of intra-taxon variability.

1.	Length 10 to 15 mm	2
	Length 15 to 20 mm	3
	Length 20 to 25 mm	5
	Length 25 to 30 mm	**New species B**
2(1).	Pronotum lateral edges convex before posterior angles; pronotum surface dull	***C. gagatina***
	Pronotum lateral edges straight or concave before posterior angles; pronotum surface iridescent	***C. bicolor***
3(1).	Head black; pronotum black	***C. gagatina***
	Head reddish; pronotum reddish	4
	Head yellowish brown; pronotum yellowish brown	**New species A**
4(3).	Antennae with 9 segments; pronotum lateral edges straight or concave before posterior angles	***C. electa***
	Antennae with 8 segments; pronotum lateral edges convex before posterior angles	***C. fimbriata***
5(1).	Pronotum surface shining; maxillary palp ratio of length to width up to 2	***C. fimbricollis***
	Pronotum surface iridescent; maxillary palp ratio of length to width 2 to 4	***C. gymnopyga***
	Pronotum surface dull; maxillary palp ratio of length to width 4 to 6	***C. erythrocephala***

Fig. 5a. The key in Fig. 4a, printed in the conventional bracketed style.

References

Dallwitz, M.J. 1980. A general system for coding taxonomic descriptions. Taxon 29, 41–46. Also available at www.delta-intkey.com/www/dallwitz-1980.htm.

Dallwitz, M.J., Paine, T.A., and Zurcher, E.J. 1993 onwards. User's guide to the DELTA System: a general system for processing taxonomic descriptions. www.delta-intkey.com/www/uguide.htm. Includes documentation of the Key program.