Running Head: Analysis of U-value

An Analysis of U-Value as a Measure of Variability

Abstract

The variability in behavior has frequently been assessed using a measure known as the U-value. Of concern in the present paper were the limits and constraints on U-value as a measure of variability. The relation between the U-value and aspects of variability was examined using three sets of simulated data. Our analysis demonstrates that the U-value as a measure of variability on its own fails to capture repetitive patterns in the sequence of responding. The U-value was shown to reflect the evenness of the distributions of responses across the categories/options used; however, when the number of categories actually used by the participant differed from the total number available, the relation between U-values and the number of categories allocated with responses was shown to be non-linear. It was also shown that the same value of U can represent different levels of evenness in response distributions over categories, depending on the number of categories/options actually used. These constraints and limitations are discussed in relation to how researchers might report on behavioral variability.

Keywords

behavioral variability; measure of variability; stereotypy; U-value

Introduction

U-value is a measure of uncertainty first reported in a paper on the analysis of serial dependencies in response chains by Miller and Frick (1949). While other measures of variability exist, such as autocorrelations (Maes 2003), conditional probability (Machado 1992; Stokes 1995) and Markov Chain analysis (Machado, 1992; 1993; 1997), U-value has been most commonly used in studies to assess the level of variability in responses where frequencies of the responses or response categories can be measured (e.g., Souza et al. 2010; Denney and Neuringer 1998; Doolan and Bizo 2013; Grunow and Neuringer 2002; Hopkinson and Neuringer 2003; Maes 2003; Murray and Healy 2013; Neuringer 2002; Neuringer and Huntley 1991; Page and Neuringer 1985; Ross and Neuringer 2002; Stokes 1995; and Ward et al. 2008). One of the advantages of using U-value as a measure of variability is that within experiments, it could be used to show the effects of contingent and non-contingent reinforcement on overall response variability. Another advantage is that it is calculated based on relative frequencies of responses to available categories regardless of the reinforcement contingencies, thus it is relatively simple to compute. However, the U-value is a molar measure of variability – it shows only the overall distributions of responses, but does not reflect the order of the sequences of those responses at a molecular level (Maes 2003; Page and Neuringer 1985; Stokes 1995). Although, the fact that the U-value is not sensitive to the sequences of responses has been reported previously (e.g., Maes 2003; Neuringer 2012; Page and Neuringer 1985; and Stokes 1995) and additional measures of variability have been suggested and used (e.g., Maes 2003; Stokes 1995), little analysis has been done to examine whether factors other than distribution evenness, such as the number of categories actually being used or the total number of available categories influence the values of U; each factor alone or the interaction of the two may influence our interpretation of variability based on the values of U. It is not clear, either, what level of variability is represented by the different

values of U (e.g., is variability in responding with the same set of options always the same for a U-value of .8?).

U-value is calculated according to the following formula:

$$U - value = -\sum_{i=1}^{\beta} \frac{\alpha_i \times \log(\alpha_i)}{\log(\beta)}$$

In the formula, β equals the number of possible categories and α equals the relative frequency of category *i*. Larger U-values reflect more even use of the available response categories compared to lower U-values which reflect less even use of response categories.

Although the formula for calculating U-value across the different studies are slightly different (e.g., using log or log2; Neuringer et al., 2000; Ross and Neuringer, 2002), the calculation is always based on the relative frequencies of all available options; and the results from using the different formula are algebraically equivalent. The U-values show whether the available options have been used equally often. Most research using U-values as a measure of variability has compared variability in responses between groups or between conditions based on the U-values obtained. However, it is unclear, when comparing U-values from different experiments or conditions where different numbers of response categories are available, whether similar U values reflect similar levels of variability. For example, is the level of variability the same for two sets of responses resulting in the same value of U? Moreover, it is also unclear whether there is a linear relation between U-values and the degree of variability; if the relation is non-linear, the degree of changes in the values of U does not necessarily reflect the amount changes in variability. For example, does a U of 0.45 reflect half the variability of a U of .9?

This paper examines U-values as a measure of variability in more detail by using three sets of simulated data; each set attempted to answer a different question about U-value. Analysis 1 consisted of 100 responses distributed over four categories; it looked at how stereotypical responding could result in high U-values. Analysis 2 consisted of 300 responses distributed over 16 categories in situations where different number of categories were used. This analysis attempted to understand how using different number of categories would impact the values of U. Analysis 3 consisted of 120 responses allocated to a number of categories, ranging from 4 to 30, available in different situations where different number of categories had been used (4, 6, 8, 10, & 12). This analysis was designed to assess how having different total number of categories impact the values of U.

Analysis 1 – Stereotypical responding patterns

The first analysis was designed to explore ways that stereotypical responding could result in high U-values. One hundred responses were distributed over four categories in four different ways; Random, Repeat, Cycle and One.

For the first type of distribution, the 100 responses were distributed randomly over four categories so that each response has equal chance of falling into any category (Random). To achieve this, the computer was programmed to pick a number between 1 and 4, randomly, 100 times. For the second type, 100 responses were distributed over four categories evenly such that that 25 responses were entered into one category before moving to another category (e.g., 111111…222222…333333…444444…; Repeat); for the third type, 100 responses were again evenly distributed over four categories, but this time in a cycling pattern (e.g., 1234123412341234…; Cycle); and for the last type, all 100 responses were allocated to only one category and no responses were allocated to the other three (denoted as type “One”).

Insert Figure 1 about here.

Figure 1 shows U-values calculated based for these four response patterns. Also shown are the frequencies of responses in each category for the four types of analysis. As can be seen, U-values were maximum (U = 1.0) for the *Repeat* and *Cycle* response patterns and minimum (U = 0) for the *One* response pattern when only one category was used. For the *Random* response pattern, frequencies of responses to each category were not strictly even because random data approximates but does not necessarily equal the intended distribution; and a U-value of .98 was obtained. Another 10 sets of 100 responses were generated randomly over four categories and the resulting U-values ranged from .95 to .99 (not shown in Figure 1).

Both of the response patterns, Repeat and Cycle, ensure equal number of responses to all four categories which result in a high U-value of 1. However, both response patterns are highly stereotypical, which is not what one might intuit from a high U-value and thus represent a limit on its utility as an index of variability. U is an effective measure of the evenness of the response allocation across categories; but it does not capture anything of the sequential relationships of responses to those categories. Sequential dependent measures of variability, such as autocorrelations (Maes 2003), conditional probability (Machado 1992; Stokes 1995) and Markov chain analysis (Machado 1992, 1993, 1997) have been used to effectively identify possible stereotypy in responses that is not discernible from differences in U-values.

Studies in which variability of behavior is reinforced have sometimes based their reinforcement schedule on the relative frequencies in relation to previous responses (e.g., Ross and Neuringer, 2002). Changes in the way reinforcement is scheduled could be made so that such stereotypical response pattern would not occur. Take the four categories in Analysis 1, for example, 16 new categories based on the original category 1, 2, 3, 4 can be created as 11, 12, 13, 14, 21, 22, 23, 24, 31, 32, 33, 34, 41, 42, 43, and 44. Now reinforcement could be

based on the occurrences of these 2-response sequences. Providing reinforcement based on the previous use of these categories could possibly prevent the stereotypical responding; high U-values would then be less likely resulted from stereotypical response patterns shown in Analysis 1.

Analysis 2 – U-values from using different number of available options

The second analysis consisted of allocating 300 responses over 16 possible categories; each data set differed in the number of categories out of the 16 to which responses were allocated. The number of categories to which responses were allocated ranged from 1 to 16. For example, if four of the 16 categories were to be used, each of these four categories were allocated 75 responses and the other 12 categories were allocated no responses. As it is not possible to calculate the log of 0, where a category was allocated no responses, log 0 was replaced by 0 in the calculation. Another approach would be to allocate a very small number to these categories, such as .00001 to make the log computable. This was also tried but made little difference to the final results. Note that responses to the used categories were equally distributed so that the maximum U- values could be obtained for the number of categories allocated with responses.

U-values were calculated for the different response patterns and are plotted in Figure 2. Figure 2 shows the relationship between the U-values (y-axis) and the number of categories to which responses were allocated (x-axis). The curve shows the maximum U-values possible if the 300 responses were evenly distributed over all the used categories for each number of categories to which responses were allocated, with no responses made to the unused categories.

Insert Figure 2 about here.

When all categories were used, 16 in this case, and when the responses were evenly distributed, the maximum U-value was 0.999 which reflects the highest variability possible for 300 responses over 16 categories. The U-value from random responses is 0.993, very close to the highest value. U-values decreased when the distributions of responses over these 16 categories were less even. For example, judging by the values of U for points *a*, *b*, and *c* in Figure 2, one could conclude that responses that resulted in the U-value for point *a* were more evenly distributed than those for point b and point *c*, with those for point *c* being the least evenly distributed. This would be true for comparisons across situations in which the same number of categories were used; the higher the U-value, the more evenly distributed responses over the categories that were used.

When only half of the available categories were used, eight in this case, even if the responses were evenly distributed (maximum variability for eight categories being used), the highest value U is 0.753 (point *h* in Figure 2). A U-value of 0.753 can also result from using more than eight categories (e.g. 9, 10, 11… 15 and 16). However, while a U-value of 0.753 is the maximum possible when eight categories are used, it is not the maximum value for 10 (point *g*), 12 (point *f*), 14 (point *e*) or 16 (point *b*) categories. This makes using U to compare variability of responses difficult as it appears that the same value of U can result from different levels of variability, depending on the number of categories to which responses were allocated. Therefore, it is worth noting that before making comparisons between the behavior of groups or subjects based on U-values, the number of categories to which responses were allocated should also be taken into account and these should be reported as well as the U-values.

One possible way of comparing the evenness of distributions when different numbers of categories were used is to look at the distance between the obtained U-value and the maximum U-value one can obtain for using the same number of categories. For example, the

U-value for point *g* (using 10 categories) is 0.753 and the maximum U-value for using 10 categories is 0.832 (point G), thus the distance for point g to the maximum U-value is 0.080. By using the same calculation, the distance to maximum U-values from point *f* to point F (12 categories containing responses) is 0.145, 0.202 from point *e* to point E (14 categories containing responses) and 0.247 from point *b* to point B (16 categories containing responses). We already discussed that for the same number of categories with responses allocated, the greater distance from the maximum U-value, the less even the distributions of responses across categories will be. Thus we can assume that the evenness of the response distribution will be greatest for point *h* and least for point *b*, even though the number of categories with responses allocated for point *h* is 8 but the number of categories that contained responses is 16 for point *b*.

Figure 2 also shows the differences in U-values between using two, four, six, eight, 10, 12, 14, and 16 categories (*d*1, *d*2, *d*3…*d*8); these differences were calculated by subtracting the U-value obtained by using the lower number of categories from U-values obtained by using the higher number of categories. As can be seen, maximum U-values increased as the number of categories to which responses were allocated increased; however, the magnitudes of the increase were non-linear. This means that changes in the sizes of U-values did not directly relate to changes in number of categories of the 16 to which responses were allocated.

Analysis 3

In Analysis 3, 120 responses were evenly distributed across four, six, eight, 10 and 12 categories when different numbers of categories, ranging from four to 30, were available for use. Analysis 3 was carried out to examine how changes in the number of available categories affects the values of U when the number of categories to which responses were allocated was fixed. Separate sets of 120 responses were created for each of the number of categories to which responses were to be allocated. For example, when four categories were

used, U-values were calculated when four, five, six … and 30 categories were available. The same calculations were carried out for using six, eight, 10 and 12 categories. As a result, there were 27 resulting U-values when four categories were used, 25 when six categories were used, 23 when eight categories were used and so forth. These U-values are plotted in Figure 3 where the numbers of available categories are shown on x-axis and the different numbers of categories which contained responses are represented by the different series.

Insert Figure 3 about here.

As can be seen in Figure 3, when the same numbers of categories were used (each series), U-values decreased as the numbers of categories increased; and this is true for all different series. This can be predicted based on how U-value is calculated; the increase in the number of categories available results in increase in the values of $\log(\beta)$ which in turn results in smaller values of U. It should be noted that, similar to Figure 2, the decrease in U-values with the increase in the number of available categories was non-linear. For example, when four categories had responses allocated (bottom series), the decrease was greater from four available categories to around 15 available categories; while the decrease was much smaller after 15 or more available categories. In addition, when comparing the trends across series, the data path for the bottom series (four categories containing responses) was steeper than the data paths for the rest of the series; the greater the number of categories to which responses were allocated, the flatter the data path. It appears that the extent to which U-values is impacted by the number of categories available depends on the number of categories actually used in relation to the total number available. As shown in Figure 3, when four categories (4/30) were used (bottom series), the shape of the data path was different from those when six, eight, 10 and 12 categories were used; the shapes of the data paths when eight (8/30), 10 (10/30) and 12 (12/30) categories were used are very similar. It is not clear, however,

whether there is a boundary percent of categories to which responses are allocated (e.g., less than or equal to 4/30) that relates to the different shapes of data paths.

General discussion

This paper examined the extent to which the U-value, a measure of variability that is frequently used in the literature, represents different aspects of response variability. As pointed out earlier, the U-value has been commonly used to show overall variability in responses; however, little analysis has been done previously to see how U is affected by factors other than the evenness of the distributions of responses. The limitation that the U-value reflects only the overall response distributions and not the sequences of responses has long been recognised (e.g., Maes 2003; Neuringer 2012; Page and Neuringer 1985; Stokes 1995); Analysis 1 of the present analysis of U demonstrated two highly stereotypical patterns that could result in the highest U-value ($U = 1$).

Another concern in using U as a measure of response variability is that the values of U are ambiguous; the same size U-value can result from response distributions with different level of evenness over categories. Barba (2012) pointed out that the calculation of U-values based on the complete set of possible response options or categories might not be appropriate when only some of the options are used. The second analysis confirmed his concern. The same value of U, for example, .75, can be the result of distributing responses completely evenly across eight categories or less evenly across any number of categories that is greater than eight – the more categories allocated with responses, the less even the response distribution. Therefore, comparing the variability across response sets using only the values of U and without information regarding the number of options each used is meaningless. However, even if information on the number of options used was provided, which set of responses is more or less variable is difficult to determine and can fundamentally default to a

subjective classification. For example, is it more variable to use more options but less evenly, or is it more variable to use fewer options but more evenly? In addition, when there are large number of options unused (e.g., say only eight categories out of 16 are used), the maximum U-value can appear to be quite low (e.g., the U-value for distributing responses evenly over eight categories out of 16 was below .8). This becomes a concern when there is a large number of possible options without enough opportunities to produce them, particularly when the number of options is greater than the number of total response required. For example, if one were to require someone to produce sequence of eight responses over three keys on a computer keyboard there would be 6561 possible sequences. If 3000 responses were required, this is quite a large number to ask of human participants, then, even if all 3000 sequences were different, the U-value would be very low.

Results from Analysis 3 shows that it is not only the number of categories to which responses are allocated that impacts the values of U but it is also the percent of categories actually used in relation to the total number of categories available. When four categories are used, changes in the values of U can be very different depending on the total number of categories available for use. For example, the extent to which U-values decrease would depend on the total number of categories available.

Summary

Analysis 1 showed stereotypical ways of responding that resulted in extremely high U-values; Analysis 2 and 3 showed comparisons of U-values need to take into account the number of the available options that were used. While the U-value might provide useful information about overall variability in responses, analyses of U-values over the response patterns revealed that it is also a limited measure of variability. First, it does not capture stereotypical patterns in responses; second, it is ambiguous as there are no specific values

distinguishing between high and low variability; and third, it is ambiguous because same values of U could result from different level of evenness in response distributions, depending on the number of options used. For these reasons, the U-value should be used with caution as a measure of variability in responses and consideration should be given to also reporting alternative measures of variability.

We suggest that future investigations of behavioral variability intending to use U-value as a measure of variability should use additional measures that capture other aspects of behavior beyond the relative frequencies of category use, in particular measures that give information about sequential relations in the data. Alternatively, reinforcement schedules that reduce the likelihood of stereotypical responding should be used. Also, when comparing the variability in the behavior of individuals based on U-values, information in regard to the number of options used should also be included (e.g., the number of options used and the frequencies of each option used). The distances of the obtained U-values to the maximum U-values possible when using a certain number of options can be calculated and these would be informative in showing the evenness of distributions when different numbers of options are used. Finally, when a task has a large number of possible options but not a large enough number of response opportunities, then the U-value is not a particularly informative measure of behavioral variability.

Compliance with Ethical Standards:

Funding: The analysis reported in the submitted manuscript forms part of the first author's PhD thesis; no funding was received.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Barba, L. S. (2012). Operant variability: a conceptual analysis. *The Behavior Analyst, 35*, 213-227.

Denney, J., & Neuringer, A. (1998). Behavioral variability is controlled by discriminative stimuli. *Animal Learning & Behavior, 26*, 154-162.

Doolan, K. E., & Bizo, L. A. (2013). Reinforced behavioral variability in humans. *The Psychological Record, 63*, 725-734.

Grunow, A., & Neuringer, A. (2002). Learning to vary and varying to learn. *Psychonomic Bulletin & Review, 9*, 250-258.

Hopkinson, J., & Neuringer, A. (2003). Modifying behavioral variability in moderately depressed students. *Behavior Modification, 27*, 251-264.

Machado, A. (1992). Behavioral variability and frequency-dependent selection. *Journal of the Experimental Analysis of Behavior, 58*, 241-263.

Machado, A. (1993). Learning variable and stereotypical sequences of responses: Some data and a new model. *Behavioural Processes, 30*, 103-130.

Machado, A. (1997). Increasing the variability of response sequences in pigeons by adjusting the frequency of switching between two keys. *Journal of the Experimental Analysis of Behavior, 68*, 1-25.

Maes, J. H. R. (2003). Response stability and variability induced in humans by different feedback contingencies. *Learning & Behavior, 31*, 332-348.

Miller, G. A., & Frick, F. C. (1949). Statistical behavioristics and sequences of responses. *Psychological Review, 56*, 311-324.

Murray, C., & Healy, O. (2013). Increasing response variability in children with autism spectrum disorder using lag schedules of reinforcement. *Research in Autism Spectrum Disorders, 7*, 1481-1488.

Neuringer, A. (2002). Operant variability: Evidence, functions, and theory. *Psychonomic Bulletin & Review, 9*, 672-705.

Neuringer, A. (2012). Reinforcement and induction of operant variability. *The Behavior Analyst, 35*, 229-235.

Neuringer, A., & Huntley, R. W. (1991). Reinforced variability in rats: effects of gender, age and contingency. *Physiology & Behavior, 51*, 145-149.

Page, S., & Neuringer, A. (1985). Variability is an operant. *Journal of Experimental Psychology: Animal Behavior Processes, 11*, 429-452.

Ross, C., & Neuringer, A. (2002). Reinforcement of variations and repetitions along three indepedent response dimensions. *Behavioural Processes, 57*, 199-209.

Souza, A. S., Abreu-Rodrigues, J., & Baumann, A. A. (2010). History effects on induced and operant variability. *Learning & Behavior, 38*, 426-437.

Stokes, P. D. (1995). Learned variability. *Animal Learning & Behavior, 23*, 164-176.

Ward, R. D., Kynaston, A. D., Bailey, E. M., & Odum, A. (2008). Discriminative control of variability: Effects of successive stimulus reversals. *Behavioural Processes, 78*, 17-24.