

## A Proposed Measure of Discussion Activity in Threaded Discussion Spaces v0.9

David Wiley, PhD  
Department of Instructional Technology  
Utah State University  
[david.wiley@usu.edu](mailto:david.wiley@usu.edu)  
<http://wiley.ed.usu.edu/>

### Introduction

For a number of reasons, there is significant interest in having our students engage in discussion in our online courses. From pragmatic matters such as the degree to which social interaction lowers student drop out rates, to pedagogical considerations around the depth of understanding students gain by negotiating the collaborative solution of problems, to simple increases in student satisfaction with online courses due to opportunities for socialization, encouraging dialogue among our students increases learning in a variety of domains and meta-domains (such as ability to work as a team). It therefore becomes desirable to be able to obtain standard measures of discussion activity within online environments. In other words, it would be useful to have the ability to ask “Are students engaging in discussion?” and “How much dialog are they engaging in, comparatively speaking?”

Computing such a measure requires an operationalization of “discussion.” At a minimum, I consider discussion to require one person to communicate to at least one other person, and for that second person to communicate back to the first. In other words, the smallest message structure in a threaded discussion space that should be labeled a discussion would be comprised of a message and a response, and would probably be rendered by a browser or other viewer like this:

- [Why did the Civil War happen?](#)
  - [Re: Why did the Civil War happen?](#)

### Operationalizing Discussion

This is, of course, a minimal example of discussion. We would ideally like our students to engage in deeper discussions than simple question answer exchanges. Consider the following two sub-threads each including five posts.

#### Sub-thread One

- [Why did the Civil War happen?](#)
  - [Re: Why did the Civil War happen?](#)
- [How has the geography of the East changed since the Civil War?](#)
- [If the South had won...](#)
- [Good websites about the Civil War](#)

Working Draft.

For the most recent copy of this paper, please contact [david.wiley@usu.edu](mailto:david.wiley@usu.edu).

### Sub-thread Two

- [Why did the Civil War happen?](#)
  - [Re: Why did the Civil War happen?](#)
  - [The role of slavery in the war](#)
    - [Who could honestly defend the practice of slavery?](#)
    - [States should be able to control their own economic destiny](#)

Sub-thread one contains five posts, as does sub-thread two. All of the posts in each sub-thread are on topic, and look like they contain information relevant to the topic. For all the similarities of these sub-threads, there is one significant difference: their structure. In sub-thread one, there are four messages and a single reply. In sub-thread two, there is one message, two replies, and two second-level replies. This structure reveals much about the degree to which students in the course are talking to each other.

In sub-thread one, many students are talking, but only one has responded. The operationalization of discussion given above requires a second person to respond to the first for discussion to occur. In sub-thread one, however, we see that the majority of the posts are monological. There could be a number of reasons for the lack of responses: top-level messages may not be amenable to responses due to their topic or voice (Wertsch, 1991), the topic of the message may not be interesting to other students for any number of reasons, students may not know the answer to questions posed in the message, etc. However, the reason why certain posts do not generate responses is outside the scope of arriving at a measure of discussion.

In sub-thread two, many students are talking, and most of them are responding to each other. It would therefore be desirable for any measure of discussion activity to assign a higher score to sub-thread two than sub-thread one.

I propose an operationalization of discussion activity suitable for computation and comparison based on the reply structure inherent in threaded discussions. I suggest that replies are strong indicators (although certainly not guarantees) that discussion is happening, and that as the level or depth of replies increases (this is easily seen in the presentation of the posts by the browser as replies move further and further to the right), the depth of discussion increases (no claim is made as to proportionality of increase).

### A Crude Measure of Discussion Activity

If reply depth can serve as an indicator of discussion activity, then one manner in which aggregate discussion activity could be measured for a group of posts would be the *mean reply depth* (MRD) of the group of messages. The issue of how many posts constitute a group suitable for analysis is left to the individual researcher, although circumstances will frequently suggest sensible groupings (e.g., one might group posts by course topic (unit on the Civil War) or time period (one semester)). The MRD for a given group of posts would be computed as shown in Equation 1.

$$d_{crude} = \frac{\sum_{i=1}^n r_i}{n} \quad \text{Equation 1.}$$

Where  $d_{crude}$  is the mean reply depth (MRD) for the group of messages,  $r$  is the reply depth of the  $i^{\text{th}}$  message, and  $n$  is the total number of messages in the group. The value of  $r$  is determined as follows: top-level messages are assigned a value of 0, first-level replies are assigned a depth value of 1 (normally rendered as indented to the right 1 indent from the left margin), second-level replies are assigned a depth value of 2 (normally rendered as indented to the right 2 indent from the left margin), etc. The value of  $d$  has a lower bound of 0 (meaning the group of messages has  $n$  top-level messages and no responses whatsoever), indicating no discussion, and no theoretical upper bound.

### A Better Measure

This approach to calculating discussion activity has three weaknesses. First, this is a measure of discussion *activity*, not a measure of discussion *quality*. A measure of quality would almost certainly require human raters to review the contents of individual messages, and would therefore require significant time and effort. Calculating the activity measure presented here can be automated to provide a quick view of the degree to which individuals are talking to each other. However, the receiver of such data *must* remember that the measure is one of activity, and not necessarily one of quality.

The second weakness of  $d_{crude}$  is that participants in online discussion groups do not always thread their contributions correctly. Webboards that structure the reply process can make threading more reliable, but (especially in mailing list archives) this problem can be significant. The problem will be reflected in the calculation as an inaccurate value in the numerator, as posts that were actually replies are counted as top-level messages (“run away children”) and top-level messages are counted as replies (“adopted children”) due to threading errors.

Several points are worth making regarding this second weakness. First, as run away children artificially deflate the value of  $d_{crude}$  and adopted children artificially inflate the value of  $d$ , it is likely that these effects will come close to canceling each other out. Second, it is my experience that, if anything, run away children are more common than adopted, meaning that  $d_{crude}$  may systematically underestimate the “true” level of discussion activity in a group. And finally, when it is desirable to achieve greater accuracy, one may adjust the numerator of Equation 1 to account for this error.

To do this, I recommend a five step process:

1. Sample a portion of the total discussion archive and calculate a sample  $d_{crude}$  (using  $n-1$  in the denominator).

2. Manually check each message in the sample and rethread run-aways and adopted children appropriately.
3. Calculate a corrected sample  $d_{crude}$  using the new threading structure (again using  $n-1$  in the denominator).
4. Calculate a correction for misthreading value  $m$  as per Equation 2.
5. Multiply  $d_{crude} \times m$  to get a mean reply depth value corrected for problems due to misthreading.

$$m = \frac{d_{correctedsample}}{d_{sample}} \quad \text{Equation 2.}$$

While this process can increase the accuracy of  $d_{crude}$  it also consumes a significant amount of human time in sampling, rethreading, and running new calculations. Because I believe the research value of the mean reply depth increases according to the extent that its calculation can be automated, I will assume that  $c$  is generally sufficiently close to 1 as to provide little additional value to the calculation. However, even though underestimations due to  $c$  are systematic, when comparing levels of discussion activity across media (such as comparing a web board to a mailing list),  $c$  should be used because the value of  $c$  likely varies across media.

The third weakness with  $d_{crude}$  comes from the treatment of top-level messages. Because top-level messages are assigned a level depth of 0, a group of messages with 2 top-level messages and 5 first level replies would have the same  $d_{crude}$  value as a group of messages with 500 top-level messages and 5 first-level replies. While technically the same amount of discussion is occurring within both groups, it would be desirable for a measure of discussion to account for this difference. In other words, the measure should reflect the signal-to-noise ratio of the group – “what proportion of the postings to this group are actually parts of a discussion?” I propose an *adjusted mean reply depth* calculated by assigning a penalty to each top-level message that has no replies. The adjusted mean reply depth calculation is shown in Equation 3.

$$d = d_{crude} \times ((n-b)/n) \quad \text{Equation 3.}$$

Where  $d$  is the adjusted mean reply depth,  $b$  is the number of top-level messages that have no replies (messages that are barren or without children), and  $n$  is the total number of messages.

### Characteristics of $d$

$d$  has a lower bound of 0 and no theoretical upper bound. To promote understanding of the calculation’s value better, some sample thread structures and associated  $d$  values are presented in Table 1, and an interpretation of some  $d$  value ranges is presented in Table 2.

Table 1. Sample thread structures and d values

<b>Thread structure</b>	A	A	A	A	A
	A	A	--B	--B	--B
	A	--B	----C	----C	----C
	--B	--B	--B	--B	-----D
	--B	--B	--B	----C	--B
<b>d value</b>	0.24	0.48	1	1.2	1.4

As an example of how the measure is to be calculated, the calculations in the first and final examples in Table 1 are described below.

*Example 1 from Table 1 (0.24)*

First, the crude mean reply depth is calculated by summing the reply depths of the five messages (0 points for each A (total of 0 points), 1 point for each B (total of 2 points)). Thus the crude mean reply depth is  $2/5 = 0.4$ . The penalty for childless top-level messages is the ratio of the total number of messages (total of 5) minus the top-level messages without children (As without Bs, total of 2) to the total number of messages (5), or  $(5-2)/5 = 0.6$ . The adjusted mean reply depth (assuming no error due to misthreading) is  $0.4 \times 0.6 = 0.24$ .

*Example 5 from Table 1 (1.4)*

First, the crude mean reply depth is calculated by summing the reply depths of the five messages (0 points for each A (total of 0 points), 1 point for each B (total of 2 points), 2 point for each C (total of 2 points), and 3 points for each D (total of 3 points)). Thus the crude mean reply depth is  $7/5 = 1.4$ . The penalty for childless top-level messages is the ratio of the total number of messages (total of 5) minus the top-level messages without children (As without Bs, total of 0) to the total number of messages (5), or  $(5-0)/5 = 1$ . The adjusted mean reply depth (assuming no error due to misthreading) is  $1.4 \times 1 = 1.4$ .

Table 2. Initial interpretations of d value ranges

<b>d value</b>	<b>Possible interpretation</b>
0 to 0.3	Monologue or lecture; no discussion
0.3 to 1.2	Simple Q & A; chit-chat
1.2 and higher	Discussion, Multilogue

Example analyses

Traffic on two discussion lists was analyzed using d and compared. Linux-Security-Module (<http://mail.wirex.com/pipermail/linux-security-module/>). “is a forum to

design, implement, and maintain suitable enhancements to the LKM to support a reasonable set of security enhancement packages.” The second list, Mailman-Users (<http://mail.python.org/pipermail/mailman-users/>) is a forum for “users and other parties interested in the Mailman mailing list management system.” Table 3 shows  $d$  values by month for the first half of 2002, and average values for the six month period.

Table 3.  $d$  values for two mailing lists

	<b>L-S-M</b>	<b>M-U</b>
Jan 2002	1.92	0.41
Feb 2002	1.40	0.52
Mar 2002	1.70	0.45
Apr 2002	1.42	0.37
May 2002	0.78	0.52
Jun 2002	1.15	0.37
Jul 2002	1.48	0.35
<b>Avg</b>	1.41	0.43

Although both these lists revolve around information technology, and more specifically, software, discussion levels on the two sites are extremely different. Mailman-Users’ discussion level would be classified as “Simple Q & A” according to the metrics shown in Table 2. This seems to make sense because the list is comprised mainly of announcements or questions from users about how a particular feature works. Neither of these types of messages generate more than a few, if any, responses.

A much higher level of discussion occurs on Linux-Security-Modules, because the purpose of the mailing list is significantly different. L-S-M is a group of physically dispersed people engaging in a collaborative design and problem solving process. Unsurprisingly, the complexity of the issues involved in designing, coding and maintaining software requires a significantly greater level of discussion than simply using finished software. In this case, the  $d$  values match our intuition regarding the performance of the measure in assigning values to amounts of discussion activity.

#### Instructional design implications of $d$

An observation may be made examining the values of  $d$  and the environmental conditions in the two groups discussed above. First, in the absence of external pressure (like homework assignments) to discuss topics with others online, deeper discussions occurred in the group dedicated to the more complex topic. Extending this observation into a design recommendation:

- When requiring students to engage in online discussions, use a complex topic as the centerpiece of the discussion.

Secondly, the structure of the  $d$  measure has implications for online discussion assignments. Requirements for posting to a course list or message board frequently take

Working Draft.

For the most recent copy of this paper, please contact david.wiley@usu.edu.

the form of “x posts per unit time,” such as “1 post per week.” However, as demonstrated above, it is not posts, but more specifically *replies* that are the critical ingredient in online discussion. An appropriate design recommendation would take the form:

- When requiring students to engage in online discussions, require a ratio of at least 4 replies for every parent post, for example, “post a question or topic beginning a new thread only once per month, and post a reply in an existing thread at least once each week.”

This assignment structure should guarantee  $d$  values in the 1.2 and higher range, and facilitate students actually engaging each other in conversation instead of firing monological shots in the dark.

### Conclusion

I have recommended a method of quickly quantifying the level of discussion activity in online groups. Alone this individual measure may not seem of much value, but I believe that the ease and speed of its calculation make it a valuable contribution to the toolkit of Internet researchers. When used in conjunction with other quantitative and qualitative measures, the adjusted mean reply depth ( $d$ ) can quickly provide valuable information about one characteristic of online groups.

The PERL program used to make the calculations reported above is available online at [http://wiley.ed.usu.edu/snail/paper\\_script.txt](http://wiley.ed.usu.edu/snail/paper_script.txt)

References

Wertsch, J. V. (1991). *Voices of the Mind*. Cambridge, MA: Harvard University Press.

Working Draft.

For the most recent copy of this paper, please contact [david.wiley@usu.edu](mailto:david.wiley@usu.edu).