Journal of Usability Studies

Vol. 12, Issue 2, February 2017 pp. 56–72

Beyond Average: Weibull Analysis of Task Completion Times

Bernard Rummel
User Research Expert
SAP SE
Dietmar Hopp-Allee 16
D-69190 Walldorf
Germany
bernard.rummel@sap.com

Abstract

Weibull analysis is an established method in technical reliability analysis for describing and analyzing the lifetime of technical parts. This paper describes the approach and demonstrates its application on task completion times from small-sample usability tests. Fitting a Weibull distribution model to observed data lets the analyst estimate task completion rates for any given time, and vice versa. Model parameters can be related to aspects of technical and cognitive efficiency, as well as factors that accelerate or decelerate user performance, and therefore are new candidate metrics for quantifying user interface (UI) efficiency.

In an analysis of time data from 144 tasks from quantitative, summative usability tests, in 98.6% of cases a 3-parameter Weibull model could be fitted; individual outlier data points had to be removed only in two cases.

The methodology also affords estimating parameters when not all test participants were able to solve the task. For task completion rates over .60, the population median time estimation (when 50% of users can be expected to solve the task) from the full Weibull model can be approximated by dividing the geometric mean of successful solution times by the task completion rate. Median time estimates should however not be understood as means, but rather as the "half-life" of the task completion process.

In the Appendix of this paper there is a link to a spreadsheet calculator that I have provided so readers can perform Weibull analyses on their own.

Keywords

usability metrics, efficiency, task completion time, statistical distribution, Weibull analysis

Introduction

Task completion time is the most popular metric for user interface efficiency (Coursaris & Kim, 2011; Hornbæk, 2006; Molich et al., 2010; Sauro & Lewis, 2009). ISO standards require completion time statistics as a mandatory part of summative test reports (ISO 25062; ISO TS 20282). However, the rich body of statistical methods that has been specifically developed both in technical reliability analysis and medical survival analysis for understanding time duration data has—as yet—received little attention in the user experience field (Rummel, 2014). Weibull analysis is a notable exception. In a nutshell, the methodology consists of fitting a Weibull distribution model to the data observed and analyzing the model's parameters.

Weibull modeling of task completion time distributions has a number of distinct advantages over classical summary statistics. First, a well-fitting distribution model provides an estimation of the task completion rate at any point in time, and vice versa. Such a model can be used, for instance, in business process modeling, in ROI considerations, or to support design decisions when comparing design variants. Second, Weibull distribution parameters (discussed in detail below) can be quite informative for analyzing the respective contributions of random and non-random influences to a process. Liu, White, and Dumais (2010) applied the method successfully to better understand website dwell time, where they related distribution parameters to specific web browsing behaviors. Such parametric information is particularly helpful when users cannot be observed directly, like in web analytics, or in large-scale unmoderated online usability tests. Third, given sufficient coverage in terms of data sets where a Weibull model can be fitted to the data observed, the approach might offer new ways to parametrize user interface efficiency in a standardized way.

In this paper, I outline the pragmatic application of Weibull analysis to task completion times, as they can be observed in usability tests. The survival analysis concept of "censored data" will be applied to account for the typical situation in usability tests, that not all test participants can solve all tasks.

In order to verify the method's applicability to real-life testing data, in the empirical part of this study, I report the results from an analysis of a large set of data from summative usability tests of business software applications. I further discuss those results and draw some conclusions for usability practitioners.

What's So Special About Time?

Consider a number of usability test participants attempting to solve a given task. For those solving the task, completion times will follow a certain statistical distribution. Let's consider such distributions in detail.

Different from many other usability metrics, task completion times are rarely normally distributed (Sauro, 2011; Sauro & Lewis, 2009; for an in-depth treatment see Luce, 1986). Because time cannot be negative, zero is a natural boundary to the distribution: On the left-hand side, the distribution cannot follow the normal distribution's familiar bell shape. In addition, on the right-hand side, time duration distributions have a typical "long tail." Often, a small number of individuals take much longer than others to solve a task, sometimes multiple times as long. This skewedness is characteristic for time distributions; it reflects the way they are generated from processes in the world that may have more or less independent steps and sub-processes, each with their own demand in time. As users take different steps, make and correct mistakes, and so on, completion time will vary between them, thus generating a distribution of times.

A very common distribution of time-to-event intervals is the exponential distribution; one might say it's the equivalent of the normal distribution in the time domain. A characteristic property of processes that generate an exponential distribution is that the rate of events does not depend on the time and history of their observation, but is constant over time ("memory-less"). Radioactive decay is a commonly known example: When exactly a single atom decays is not predictable; however, a large set of atoms will decay at a constant rate. Readers will be familiar with the term *half-life*—the time until half the substance has decayed. The next half of the remaining substance (i.e., 25% of the initial mass) will take exactly the same time to decay.

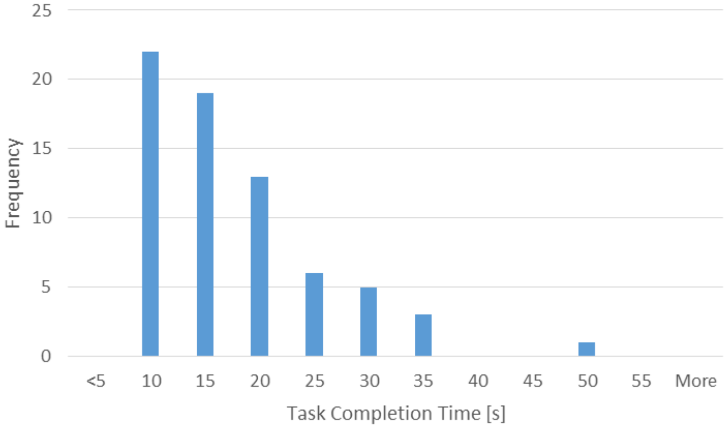

Figure 1. Distribution of task completion times in an unmoderated online usability test with 69 participants. This figure shows data in a histogram, time scale values are upper-interval boundaries (Rummel, 2014).

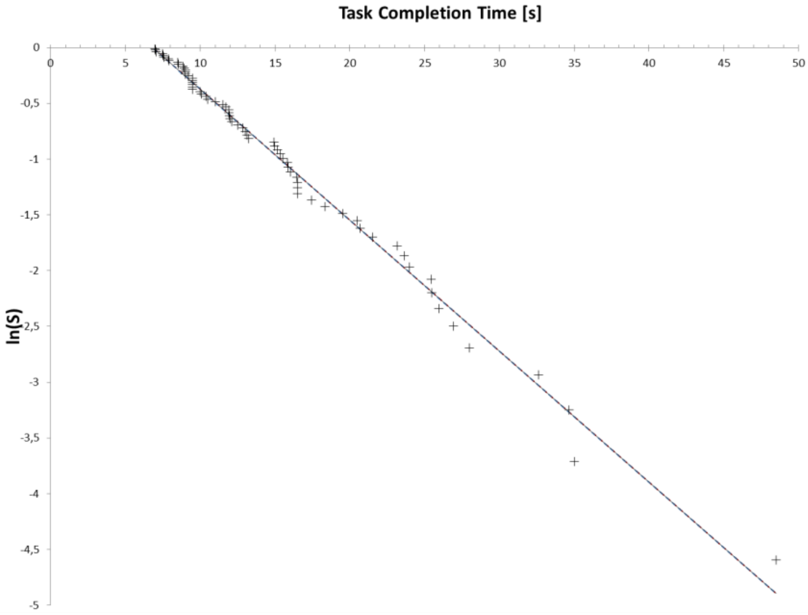

Figure 2. Distribution of task completion times in an unmoderated online usability test with 69 participants. This figure is an exponential probability plot (Rummel, 2014).

Figure 1 and Figure 2 show an example of completion times from an unmoderated online usability test task completed by 69 participants. Note the skewed shape of the histogram on the left-hand side of Figure 1; it does not at all look like the familiar bell shape of the normal distribution. Figure 2 shows, for the same task, the percentage *S* of participants still working on the task (the "survival" function) as it declines over time on a logarithmic scale. After the first participant solved the task at t_0 = 6.8s (more on t_0 below), the straight arrangement of data points indicates that from this point in time, the percentage of users solving the task in a given time interval was actually constant: Completion times for this task are exponential-distributed. This means they behave much like radioactive decay: If half the users solve the task at time t, the next remaining half (25%) will need 2t, the next remaining half (12.5%) 3t, and so on, hence the "long tail" in the histogram. Note how the singular data point around 50s in Figure 1 perfectly fits into the linear arrangement in Figure 2—it is not an outlier at all. If such a distribution can be observed, it is fair to assume that the generating process, users attempting to solve the particular task, is driven mostly by randomness[1].

Rummel (2014) observed that task completion times following an exponential distribution are rather common. The exponential distribution model however only fits when a constant offset time t_0 is subtracted from the individual times, and therefore is easily overlooked. The exponential distribution model with an offset time affords an interesting interpretation because the model mathematically splits the process into a constant (t_0) and a stochastic part (the exponential). As a constant, t_0 can be interpreted as the time contribution of all processes that add basically a constant time to the distribution: The time the system needs to respond to user input, and the amount of time a user needs to merely operate the user interface, that is, to click through the task. System response and mere operation time have typically negligible variance, so it is plausible to assume they drive the constant part of the model. The exponential-distributed part of the model, on the other hand, would result from the process of the test participant meeting with, and eventually overcoming, the various challenges imposed by the task and the user interface. The "selection" and time costs of those challenges are subject to a great deal of randomness, which may explain the exponential behavior of the time distribution. If the task solution rate for this part of the process is indeed constant, it is fair to assume a basically stochastic generating process. The half-life of this process then would be a metric of the system's *cognitive* efficiency, as opposed to t_0 describing its *mechanical* efficiency. For mathematical convenience, instead of a half-life, reliability analysts have defined the *characteristic time* τ as the time, discounting the offset, when 37% (= 1/e; e is Euler's number) are still "alive," that is, working on the task.

To summarize, the exponential distribution model separates constant *click-thru time* from stochastic *think time*. Typically, think time is considerably larger than click-thru time, which nicely illustrates the importance of usability, compared to system performance.

The Weibull Distribution

Obviously, solving a usability test task is not a purely random process, even though it may be heavily influenced by random variables. The Weibull distribution is a convenient extension of the exponential distribution model that can cover certain systematic deviations from the exponential's "memory-lessness." This made it extremely popular in reliability analysis (Tobias & Trindade, 2012). It has three parameters: offset time t_0, characteristic time τ, and the so-called shape parameter γ.

Offset and characteristic time conceptually match the parameters of the exponential distribution and can also be interpreted as "click time" t_0 and "think time" τ. The shape parameter γ describes systematic deviations from the exponential model. If $\gamma = 1$, the Weibull reduces to the exponential distribution with the same characteristic time τ. A shape parameter $\gamma < 1$ indicates that test participants, the longer they are working on a task, solve it even slower than expected by the exponential model (that is, if the process were purely random). If $\gamma > 1$, they are faster, possibly by an implicit learning process or another accelerating factor. This makes γ rather interesting when inspecting time distributions in cases where users cannot be directly observed; it points at non-random (i.e., probably causal) factors that influence user performance in a positive or negative way.

Figure 3 shows three simulated Weibull time distributions with $t_0 = 30$s, $\tau = 100$s, and varying shape factors γ. The percentage *S* of users still working declines over time, beginning at t_0. Note how for $\gamma = 1$ data points align straight, indicating an exponential distribution. For $\gamma = 1.5$, completion times are shorter that in the exponential distribution, indicating an acceleration in the process. For $\gamma = 0.8$, completion times are longer. Note also how all curves intersect at $\ln(S) = -1$, that is, $S = 1/e$: The intersection point indicates the characteristic time τ. Because all curves are translated by $t_0 = 30$s, the intersection occurs on the time scale at $t = 130$s.

[1] This doesn't necessarily mean the process is inherently chaotic. Rather, influencing variables can be randomly distributed, creating the appearance of a stochastic process. The randomness then lies in the unsystematic selection of influencing factors in the experiment.

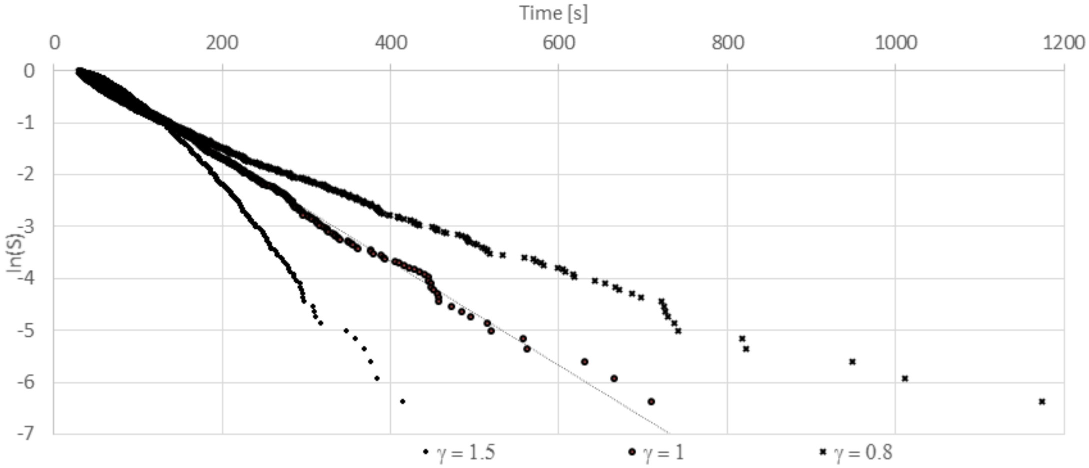

Figure 3. Simulated task completion time distributions with t_0 = 30s, τ = 100s and varying shape factors γ. See text for explanations.

To summarize, with the three parameters t_0, τ, and γ, the Weibull model—if it fits the data—has the potential to numerically describe three rather interesting concepts: click time t_0, think time τ, and acceleration γ. In addition, once parameters are known, expected task completion rates for any given time can be calculated directly from the following model equation:

Equation 1

$$TCR(t) = 1 - e^{-(\frac{t-t0}{\tau})^{\gamma}}$$

Censored Time Data

A common problem with time data is that they are not always observable for all participants in a study. For example, in a medical survival study, participants may —hopefully—survive until the end of the study, so their time of death is not observed. In usability testing, in a very similar manner, a test participant may not solve the task within the time limit, give up, or come up with a wrong solution. Here again, the actual solution time cannot be observed.

In the usability field, it is currently considered best practice to discard time data from unsuccessful test participants (e.g., Sauro & Lewis, 2012). However, researchers often do have the information available that a test participant has actively worked on the task in the first place, and even how long they tried, and may very well make use of this information. Survival analysts have developed a range of methods to deal with so-called censored data; for a comprehensive treatment see Klein and Moeschberger (2003). Rummel (2014) discussed this problem in detail, in the context of usability testing. For the purpose of this study, a slightly simplified approach was taken that will be described below together with the modeling technique.

Method

In this study, I set out to apply Weibull distribution modeling to a set of usability test data. I shall first describe the modeling technique used, then the data set.

There are various methods for fitting distribution models to time data. Tobias and Trindade (2012) identified two different general approaches: maximum likelihood estimates (MLE) and linear rectification, also known as probability plotting. MLE generally have higher accuracy and robustness than other methods, in particular when a large number of data is available for analysis. In usability tests, however, sample sizes above 15 are rather rare. In addition, it is rather common that not all test participants can solve a task, and data sets can include outliers that have to be identified and dealt with before parameters can be meaningfully estimated. The second approach, linear rectification or probability plotting, affords dealing with these issues in a pragmatic and effective manner.

The basic idea of probability plotting is to reformulate the distribution's model equation (such as in Equation 1 in The Weibull Distribution section) in such a way that a linear equation emerges, which then can be approximated with simple regression techniques. For the exponential distribution, logarithmizing the model equation suffices; Figures 2 and 3, which depict ln(S) over time, are actual probability plots for the exponential distribution. For the Weibull distribution, more complex transformations (described below) are necessary to "linearize" the model. Linear relationships can be easily visualized and verified in scatter plots. When the plot's axes are scaled appropriately to reflect the transformations used to linearize the respective distribution's model equation, data points align along a straight line, if—and only if—the distribution model is applicable to the data (note how in Figure 3, only the data points for γ=1, where the Weibull distribution is in fact exponential, align along a straight line). Regression parameters then can be used to calculate distribution parameter estimates and to assess the goodness of fit of the model. This technique affords verifying distribution models, identifying outliers, and estimating parameters in one analysis step. A small number of data points obviously limits its accuracy, but not its applicability; as long as a regression line can be drawn, the method yields results.

Rummel (2014) described in detail how to use the probability plotting method for analyzing task completion times with regard to several distributions, and provided a spreadsheet to perform the necessary calculations. In this study, I apply this methodology to a set of task completion times collected in a number of usability tests conducted at SAP in 2012-2015. In the Appendix in this paper, I have provided a link to a specialized spreadsheet for Weibull probability plots. I shall first briefly describe the probability plotting method, then the data set.

Weibull Modeling with Probability Plots

A probability plot is a scatterplot of observed task completion times against their corresponding percentiles in the population. These percentiles form the survival function *S* and need to be estimated. This can be done by ranking task completion times by magnitude; dividing rank numbers by sample size would yield a crude estimator of *S*. In order to achieve a more accurate estimation, we need to account for "censored″ data points. Suppose a test participant failed a task at a certain time. Up to that time they have worked on the task all right so their percentile rank in the population cannot be higher than that of someone who solved the task within this time. We don't know the time when they eventually, hypothetically, would have solved the task—this information is "censored." In the test paradigm that I used in this study, censoring, however, typically doesn't happen at random. (For details on random vs. non-random censoring see Rummel, 2014. In fact, random task failure—for instance, due to unprovoked equipment breakdown—is so rare that for this study I chose to treat all failure cases as non-random.) It is fair to assume that any test participant who gives up, or gives a wrong solution to a task without noticing their mistake, would need at least as long to eventually solve the task as the slowest successful participant. With this assumption, we can assign unsuccessful participants a rank (and survival percentile) behind this participant. The exact time and percentile don't matter here, for only times of successful test participants are used in the probability plot. For the purpose of this study, the survival function estimate that I used for plotting is calculated using the modified Kaplan-Meier (K-M) Product Limit recommended by the National Institute of Standards and Technology (NIST/SEMATECH, 2012a; see also Tobias & Trindade, 2012) for small samples. This estimate, in case of 100% task completion rate, converges to the uncensored case.

Once times and survival function values are known, a scatterplot can be drawn. The axes of the plot are scaled specifically for the distribution under consideration. For the Weibull distribution, the time axis is set up as the horizontal axis in natural logarithmic scale (ln[time in seconds]). The vertical axis shows the double logarithm of the survival function, ln[ln(1/S)]. If the Weibull model fits, data points in the scatterplot will align along a straight line (Figure 4A). Outliers can be immediately spotted by their deviation from this line by more than random fluctuations; systematic deviations of data points from the straight line indicate that a different distribution model might be more appropriate, or there are incoherencies in the process.

By fitting a regression line to the scatterplot, distribution parameters can be estimated. The regression's goodness-of-fit parameter R^2 is a metric for the distribution model's goodness-of-fit. If the fit is good, the regression line crosses the (ln) time axis at τ and has a slope γ.

In a Weibull probability plot, the offset time t_0 needs to be subtracted from the data *before* they are entered into the plotting algorithm. Unfortunately, the value of t_0 initially is unknown. Conceptually, t_0 must lie in the interval between 0 (there is no such thing as negative time) and the minimum observed task completion time (fastest participants being as fast as an expert who "doesn't have to think"). Within this restriction, we can choose t_0 to maximize the model fit, namely, the regression equation's R^2 value. Using Microsoft Excel's Solver Add-In, this can be automated in a spreadsheet by defining a cell containing a t_0 value. Using the RSQ function, calculate R^2 from the x and y columns of the scatterplot that are, respectively, ln(Observed Time – t_0) and ln[ln(1/S)]. By indicating the cell containing R^2 as the criterion to be maximized, and the cell containing t_0 as the cell to be manipulated, Solver quickly converges[2] to a t_0 value that maximizes the model fit R^2.

The resulting regression equation's R^2 parameter can be used to assess the distribution model's fit to the observed data. R^2 describes the percentage of variance explained by the regression model. A significance test can be based on a table of critical values provided by Filliben (1975; after NIST, 2012b) for various sample sizes. On the p = .05 level, for 18 participants, one would have to reject a distribution model if $R^2 < .89$. For the purpose of this study, with its varying sample sizes, let's choose a more conservative critical $R^2 = .90$.

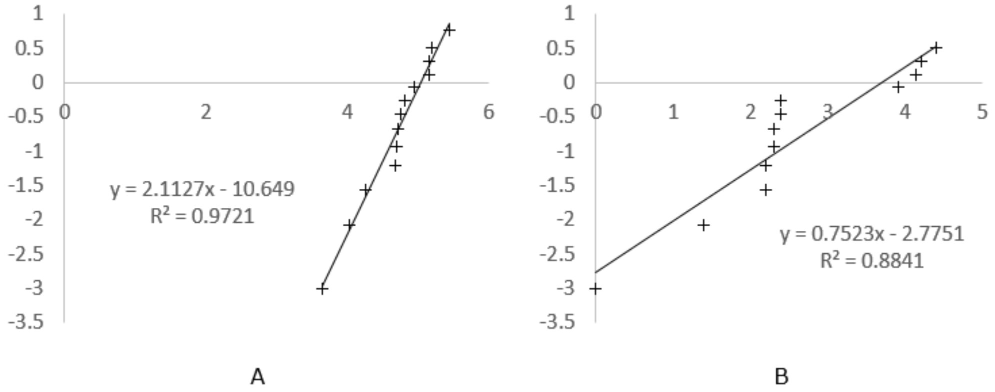

Figure 4. Two Weibull probability plots with good (A) and not-so-good (B) model fit. Vertical axis is ln[ln(1/S)], horizontal axis ln(time – t_0); time in seconds. Note the lower R^2 value and the "edges" in the data point arrangement in B. The underlying process incoherence in this test was caused by a spelling-sensitive search function.

Data Set

Task completion times that I used in this study come from a series of summative usability tests, conducted at SAP between 2012 and 2015, on business applications in various states of technical and design maturity. It should be noted that this application sample is of course not representative of SAP's product palette; applications were chosen by "test-readiness" and various feasibility criteria rather than market readiness. All tests used in this study were

[2] The Solver algorithm does not always converge to a value between 0 and t_{min}, depending on the starting conditions. Best results have been achieved with $t_{0,\ initial} = t_{min}-1$; in a few cases manual maximization of R^2 was necessary.

conducted following a strict protocol that was identical throughout the series. After some warm-up tasks, participants were given written task descriptions and were asked to paraphrase to the test moderator to ensure proper task understanding. In desktop system tests, participants used a separate monitor, keyboard, and mouse connected to the test notebook. With this setup, the moderator could set up defined starting conditions on the notebook before switching the display to the participant's monitor, which signaled the beginning of the task. Task time was measured manually from the appearance of the task screen to the participants' declaring either their solution to the task or failure. For the mobile device tests, time was measured from handing the device to the participant to their handing it back after they felt they completed or abandoned the task. Moderators gave assistance according to a strict policy: When test participants asked for help, they were first asked to re-read the task description. If this didn't suffice, they were informed (if applicable) that the solution could not be found on the current screen. Only if this didn't suffice either, minimally informative help was given (e.g., "Where haven't you looked yet?"). Further assist requests were scored as task failure.

Results

Given the intent of this study, to assess the Weibull model's applicability in usability testing, I shall first present results pertaining to the coverage of the model, that is, the extent to which Weibull models can be found to fit observed data. Once this is established, the Weibull models' respective parameter estimates can be examined in more detail.

Model Coverage

Completion times from 144 tasks out of 16 tests were selected for this study. Based on adherence to the above described protocol, absence of known artificialities and number of usable data points (a minimum of five participants solving the task that were not outliers), 129 tasks were conducted on desktop systems, 10 on smartphones, and 5 on tablets. For each test, participant numbers ranged from 8 to 18 with a median of 17. Task completion rates ranged from .22 to 1 with a median of .75.

For the Weibull model with offset time, the linear rectification model fit parameter R² exceeded the critical value of .90 in 141 cases (98.6%). Figure 5 shows a histogram of the R² values observed. Most R² values are indeed in a range where the model fit can be considered quite good; the median was R² = .966. Individual outlier times had to be removed in 2 cases (1.4%).

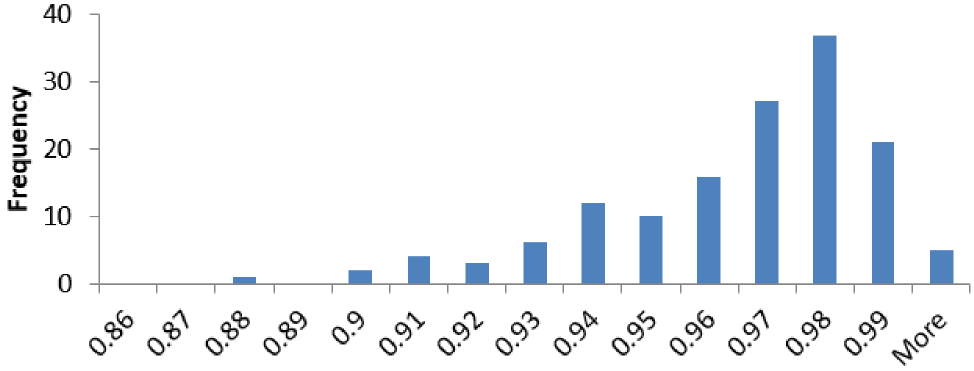

Figure 5. Frequencies of R² values indicating goodness-of-fit for Weibull distribution models of the tasks investigated. Ordinate values are upper interval bounds.

Parameter Estimates

The following section gives an overview of findings with regard to the Weibull model parameters offset time, characteristic time, and shape parameter. Because this is the first inventory of Weibull parameters in the usability testing domain, the treatment is necessarily exploratory. In addition to the Weibull parameters, I'm investigating median task completion time, a popular metric in usability research, and its relationship to Weibull models.

Offset time

As I discussed previously in this paper, the offset time t_0 can be interpreted as a constant "click time" in which test participants need to click through a task on the ideal path. For this interpretation, it is essential that this click time can not only be theoretically postulated but actually quantified in a reliable way.

Figure 6 shows a scatterplot of Weibull model estimates for t_0 by minimum observed times t_{min}. The great majority of data points align along the identity line. In a small number of cases there are very low t_0 estimates from the Weibull model; nine actually equal 0. A closer inspection of these nine low-estimate cases does not readily show anything exceptional, neither in terms of task completion rate nor model fit. However, because in the modeling technique applied here, the estimation of t_0 is based on the entire data set, so it is certainly possible that systematic influences in later data points have affected the estimate. Indeed, all but one of the nine $t_0 = 0$ cases have shape factors greater than one, indicating faster-than-expected task completion in particular in the slower test participants.

Cases where the t_0 estimate lies neither close to 0 nor the minimum observed time t_{min} are rare; Figure 6 shows only three such cases. Apparently t_{min} is a rather good approximation of t_0, as long as the latter is not 0.

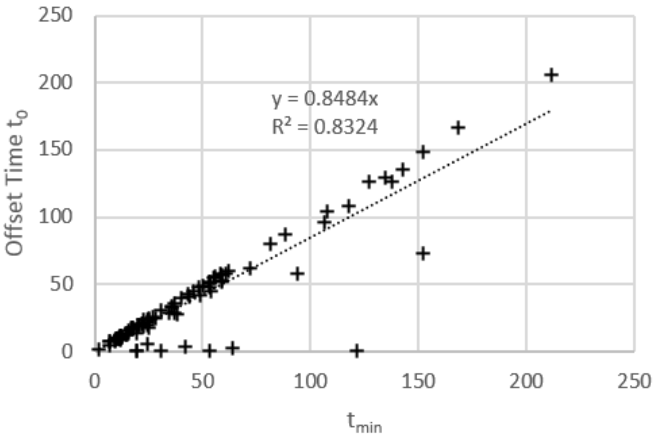

Figure 6. Scatterplot of Weibull model estimates of the offset time t_0 vs. minimum observed times for the same task, respectively.

Characteristic time

The characteristic time τ is a metric of the duration of the stochastic process overlaid to the constant offset time t_0. For exponential distributions without offset and 100% task completion rate, the characteristic time τ would equal the arithmetic mean of completion times. In usability tests, the situation typically is different: Offset times are common, as well as failed task attempts, and a shape parameter might also be present (see the Shape parameter section). In particular, the influence of the task completion rate (TCR) deserves attention. In case TCR is low, averages calculated only from successful participants can be expected to bias estimates because only the more proficient participants are considered. If, on the other hand, characteristic times estimated from the Weibull model excessively deviate from average successful times, this raises suspicion towards their validity.

In 11 cases, τ indeed is more than ten times larger than the average of successful completion times. In all those cases, the task completion rate is < 0.5. Note that τ is the time when the Weibull model predicts that 63% of test participants solve the task, which at a task completion rate of 50% obviously lies outside the actually observed range. With the number of data points also diminishing with low TCR, inaccurate estimates are to be expected.

To further investigate the relationship of task completion rates, average successful times, and characteristic times, consider Figure 7. The ratio of characteristic and average successful times (which would be 1 if both are equal) is drawn on a logarithmic scale over the task completion rate. Data points are dispersed along a band of quotient values that, when looking from right to left, moves up and widens as task completion rate decreases. In case of task completion rates > .6, there is a fairly (log) linear relationship to task completion rates that reflects the above mentioned task completion bias. Because t_0 is included in the average but not in τ, τ can be smaller than the average successful completion time. Below a task completion rate of .6, the band starts dissolving. Around TCR = .5, higher quotient values start appearing, and below .3 the band disintegrates to widely scattered values. Apparently a TCR of .6 is a critical value for interpreting τ. For lower TCR, when τ is outside the range of actually observed completion times, estimates become increasingly inaccurate.

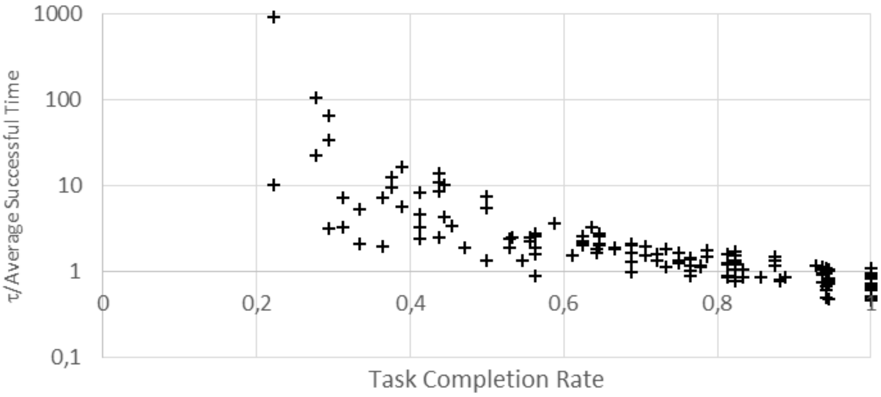

Figure 7. Scatterplot of the ratio of τ and the average successful task completion time (logarithmic) vs. task completion rate for the same task.

Shape parameter

The Weibull distribution model introduces a parameter that is new to the discussion of task completion time models, the shape parameter γ. Figure 8 shows a histogram of γ values found in the present dataset, within the 113 cases where a Weibull model could be fitted with $R^2 > .90$, and task completion rate was > .5. The distribution has its peak in the 0,8-0,9 bin, indicating that in the majority of cases, the slower test participants needed a slightly longer time to solve the task than expected "by chance" (the exponential model). This is according to expectations because immature applications were tested. However, there is also a great deal of cases with γ > 1 (32; 28%) where the opposite is true, that is, task completion was accelerated.

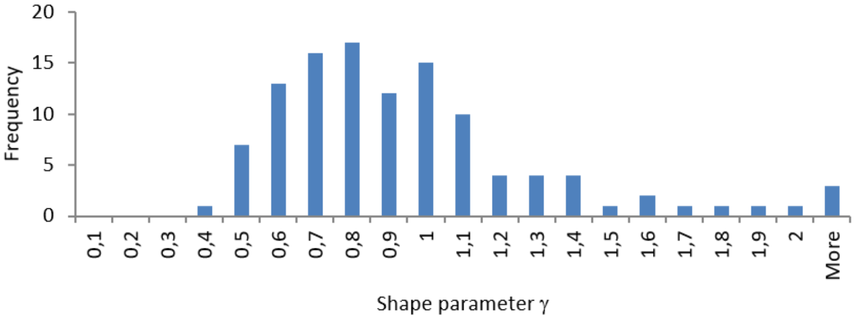

Figure 8. Frequencies of shape parameter γ values for Weibull distribution models of the tasks investigated. Ordinate values are upper interval bounds.

Previously in this paper, I suggested that a small shape factor might indicate a detrimental influence on test participants' performance. This should also be visible in other usability parameters. In fact, the correlation between shape parameters and task completion rates is r(111) = .34 (p < .01), indicating smaller shape parameters in case of lower task completion rates.

Point estimate: Median time

In many applied cases it is useful to have a single point estimate that can be benchmarked or otherwise compared to third-party research. Sauro and Lewis (2012) recommended reporting the median time, or in case of small samples, the geometric mean which is a good estimator for the population median when samples are small and the task completion rate is sufficiently high. Median time is the time when 50% of test participants solved the task. The Weibull model affords estimating a *population* median time, simply by solving the model equation for a completion time t_{50} at S = .50. This estimate is independent from the actual task completion rate because the modeling process takes task failures explicitly into account.

For task completion rates above .60, Figure 7 suggests a close and stable relationship between τ, TCR, and the average (log) time calculated from successful participants only. If this relationship is systematic enough, can we use it to estimate a population median, that is, to correct for the bias introduced by task failures?

The modified K-M censoring estimate used in the Weibull modeling process, in combination with the assumption that failing participants would need at least the solution time of the slowest successful one, basically reduces the percentage of cases used for estimation proportionally to the task completion rate. So why not estimate t_{50} by dividing the geometric mean of successful test participants' completion times by the task completion rate? Figure 9 shows a scatterplot of such estimates against estimates from the full 3-parameter Weibull model. With a restriction to tasks with completion rates > .60, the estimates correlate to r = .98; differences are in the 10% range. For lower task completion rates, the correlation diminishes rapidly (.92 for TCR $\geq$ .5; .85 for TCR $\geq$ .4).

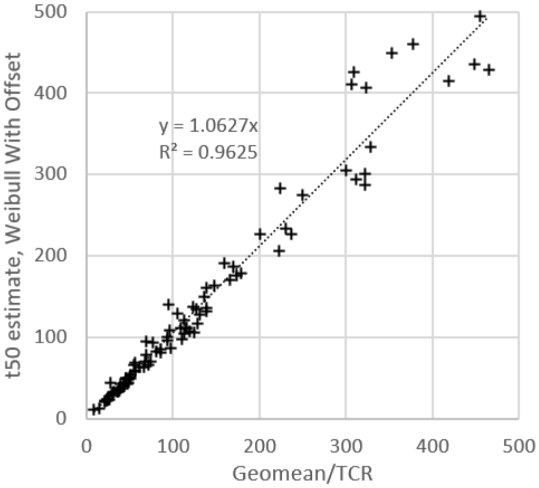

Figure 9. Scatterplot of Weibull model estimates of the population median time t_{50} vs. the geometric mean of successful task completion times, divided by task completion rate for the same task. Plot shows only TCR > .60 cases.

Discussion

In this study, I have demonstrated the applicability of the Weibull distribution model to usability test data, and have given a first orientation to which extent this model can be expected to apply. Typically, more than one distribution model can be reasonably well fitted to a set of data, so the choice of a distribution model depends to some degree on the researcher's discretion and intentions (Tobias & Trindade, 2012, detailed guidance provided on p. 95). The 3-parameter Weibull model has distinct advantages for usability practitioners. First, apparently it covers such a wide range of observed distributions that it can, for practical purposes and with reasonable precautions, claim universal applicability. Second, as a straightforward extension of the "memory-less" exponential distribution model, Weibull model parameters can be interpreted in terms of key factors in user experience research.

The offset time t_0, as a constant, can be related to the less variable components of task completion time, that is, technical performance and the time needed to click through a process on the ideal solution path. This interpretation, suggested by Rummel (2014), needs to be taken with a grain of salt. A comparison with the minimum observed time t_{min} is instructive: The fact that in most cases $t_0 \cong t_{min}$ supports its interpretation as click time; however, the equation does not always hold. The data set investigated here contains a number of instances with low or very low offset times, most of which, but again not all, show higher-than-usual shape factors γ. Consequently, if t_0 is different from t_{min}, it is difficult to tell where the difference comes from—whether t_0 or t_{min} is the actual click time. Because t_0 is numerically rather small, compared with average or characteristic times, small differences will have a relatively large impact. Consequently, quantitative interpretation of t_0 requires great caution. In practice, our experience at SAP has shown that in comparative test designs, where test conditions and process characteristics are comparable between alternatives, interpreting t_0 is certainly possible and meaningful. As a benchmark metric, where KPIs may be discussed out of context, t_0 is rather questionable unless it is close to t_{min}. Rather than determining click time from the distribution of usability test task times, practitioners may prefer considering other data sources—if they are available, which in online research, often they are not.

The characteristic time τ, as a model parameter, represents a single-point efficiency metric of the stochastic part of the process. The above discussion of t_0 shows that τ is actually the more interesting parameter, as it does not depend on the technical and mechanical efficiency of the UI and may very well, as Rummel (2014) suggested, be interpreted as cognitive efficiency. However, because it corresponds to the time (discounting t_0) when 63% of users solve a task, in case of lower task completion rates, it can be a rather theoretical figure. For task completion rates above .60, it is certainly interpretable. Between TCR's of .40 and .60, it should be interpreted with care; below .40 the metric becomes questionable (one may indeed question the meaning of a predicted time when 63% of users would solve a task, when less than 40% could demonstrably solve it).

The shape parameter γ is rather new to the discussion of user interface efficiency. The mere fact that test participants need rather long times to complete a task is not necessarily an indicator of bad usability or UI efficiency. A task may take a long time simply because it is complex. Also, the existence of test participants who need, say, twice as long "as average," is not an indicator per se because instances of long completion times are characteristic for time data as such. A small shape parameter γ, however, is indeed an indicator of usability problems, as task completion happens in an *even slower* way than a "random decay" curve would predict. The distribution of shape parameters found in this study indicates that this is common, typically to a smallish extent (most γ's are around 1), and with a number of notable exceptions where test participants were actually faster than expected just by chance (i.e., under the exponential model).

Taken together, the Weibull distribution model and its parameters have the potential to re-open the discussion "what to report" (Sauro & Lewis, 2010) with regard to task times. ISO/IEC 25062 (2006) requires that task times are reported as "mean time taken to complete each task, together with the range and standard deviation of times across participants" (clause 5.4.4.2 Efficiency). It now becomes clear that this guidance, obviously inspired by the normal Gaussian distribution model, can be misleading. Consider, as a special case of the Weibull distribution, an untranslated, "pure" exponential distribution. This distribution's standard deviation equals its arithmetic mean, and both equal characteristic time τ. The standard deviation, in the exponential and Weibull distribution family, therefore is not only a *dispersion* metric but also one of *magnitude.* The mean, on the other hand, is not at all the distribution's midpoint, nor is it independent from the standard deviation, as would be expected in the normal distribution model. Not only are the values of those parameters inappropriate descriptions of a time distribution, but the very concepts carry a serious risk of misunderstanding.

Obviously, communicating the Weibull model and the meaning of its parameters to stakeholders who are not trained in statistics is a great challenge, which may take valuable meeting time in test results communication. Sauro and Lewis (2010, 2012) pragmatically addressed these issues by recommending to logarithmize all times before calculating means, standard deviations, and confidence intervals. This effectively compensates the skewedness of the distribution and solves most practical problems in significance testing. In this study, I have validated this approach: The log transformation indeed can "normalize" a Weibull distribution to such an extent that practitioners can rely on the robustness of classical statistical methods. However, the risk remains that stakeholders keep thinking in terms of normal distribution concepts. In addition, the findings in this study suggest two important extensions to the logarithmic approach.

First, as I argued in this paper (see also Rummel, 2014), low task completion rates may distort metrics derived from only successful test participants. The empirical results of this study support this claim and further demonstrate that, as long as the task completion time is .60 or above, the effect of varying task completion rates can be effectively compensated by dividing the geomean of successful task times by the task completion rate[3]. Equation 2 yields a reasonable estimate for the (population) median completion time, which is close enough to the full Weibull model-based estimate for many practical purposes.

Equation 2

$$t_{50} \cong t_{\text{geomean,succ.}}/\text{TCR}.$$

Second, the Weibull model provides insight into the practical meaning of such median times. With most γ values around 1, most distributions are close to the exponential one. Median times here are *not* the middle, but rather the half-life of task completion. As a rule of thumb, the half-life of the stochastic task solution process part would be somewhere near median time minus minimum time (t_{50}-$t_{\min}$). Considering that $t_{\min}$ usually is much smaller than t_{50} and often practically negligible, the following Equations 3 and 4 could be used as examples to roughly estimate further quantiles.

Equation 3

$$t_{75} \cong t_{\min} + 2(t_{50}-t_{\min}) \cong 2\ t_{50}$$

Equation 4

$$t_{87.5} \cong t_{\min} + 3(t_{50}-t_{\min}) \cong 3\ t_{50}$$

This reasoning has considerable consequences, in particular for business applications: A substantial and quantifiable proportion of users can be expected to take multiple times the median time to solve a task. Ignoring the exponential distribution's long tail would instigate undue stress on workers and serious misplanning in operations management. A better understanding of time distributions would spare project managers a lot of stress and would help to remove incorrect project management myths and folklore from the world.

Readers familiar with statistical literature will have, so far, missed a discussion of confidence intervals and reliability of distribution parameters. Confidence intervals have been left out of this paper for two reasons. First, the focus of this study was to provide an overview on parameters found, not individual instances of usability test results. Second, the calculation of confidence intervals requires a different approach from the one taken here, namely, maximum likelihood estimations (MLE). I refer interested readers to the literature presented in this paper, and especially to Tobias and Trindade (2012) treatment of the subject (see pp. 116–120 in their book; they also offer readers spreadsheet tools that are available for download with their book).

[3] Analyses not reported here show that this finding is robust with regard to deviations from the lognormal distribution, where the geomean would be mathematically equal to the median.

MLE methods don't afford checking data for outliers, nor a model-based estimation of t_0. As demonstrated in this study, I show that removing outliers is rarely necessary, and t_{min} is a good initial estimator for t_0, so MLE methods can be confidently applied.

As for the reliability of Weibull parameter estimates, it cannot be estimated from the data set used in this study; the studies that I worked with had up to 18 participants—small data sets indeed for estimating three parameters. Resampling analyses would further reduce the number of usable data points. Since in the future, as more large data sets will become available from online usability studies (Albert, Tullis, & Tedesco, 2010), this question will hopefully be addressed soon.

Conclusions

The Weibull distribution model with offset time can be expected to cover a wide range of task completion times from usability tests. Once a model has been fitted to observed data, the model Equation 1 can be used to estimate expected task completion rates for any point in time, and vice versa. In addition, the Weibull model parameters provide insights for which factors contributed to the observed task completion times. The offset time t_0 reflects time consumption by technical performance and click path length. The characteristic time τ indicates the duration of the more stochastic process of UI usage, which is mostly driven by task complexity and the design quality of the user interface. The shape parameter γ indicates whether, apart from the stochastic process part, factors are present that accelerate or inhibit user performance.

A further insight from Weibull analysis is how practitioners can deal with varying task completion rates. Simple division by the task completion rate suffices to estimate population median completion times from the geometric mean of successful task completion times, with reasonable accuracy.

The Weibull model, however, clarifies that those median times are not midpoints, nor is the distribution's "long tail" a mere exaggerated measurement error. Most practitioners and stakeholders today are trained to think in terms of a normal distribution, that is, means as midpoints and standard deviations as dispersion. These concepts are misleading in the domain of task completion times. Usability practitioners should take care to correct that and point out that median times are rather "half-life" times than midpoints.

Tips for Usability Practitioners

Weibull analysis is particularly useful

- For analyzing time data from unmoderated studies where individual observation of participants is difficult. Cheaters, outliers, and other anomalies show up in probability plots as deviations from an otherwise orderly distribution model. Small shape parameters point at systemic "friction" in the UI.
- For estimating task completion rate over time, for instance, when discussing ROI of usability investments. Stakeholders are often surprised to learn how many users may take several times as long as "average," without being outliers.
- For differentiating between "click time" and "think time." The former is the one IT departments care about, the latter is the one with the long tail that slows down business processes, and where usability investments take most effect.

Of course, you'll still need qualitative observations to determine corrective design actions. The quantitative data, however, helps building business cases. For performing a Weibull analysis,

- A Microsoft® Excel™ workbook is available to perform the necessary calculations for model fitting, and calculating TCR by time, and vice versa. See the Appendix for a link to the workbook.
- Begin your analysis by inspecting the exponential probability plot for any anomalies and outliers. Investigate anomalies, and remove outliers before proceeding.
- Use Solver to find the offset time t_0 that yields the best-fitting Weibull model. If Solver doesn't converge, manually fine-tune t_0 so that the data points in the Weibull probability plot align straight. A good starting value for t_0 is t_{min} -1.

- Think of Weibull distribution parameters t_0 as "click-thru time," τ as "thinking time", γ as "acceleration." If you find $\gamma < 1$, there is some "friction" in the UI that slows users down. Investigate what the friction might be.

In any case, short of Weibull modeling,

- You can estimate the time when 50% of users can be expected to solve the task (the population median) by dividing the geometric mean of successful task completion times by the task completion rate. The division effectively compensates the effect of varying task completion rates, as long as they are > .60.
- Mind that median times are rather a "half-life" than a midpoint. Expect, and communicate to expect, much longer task completion times. Time distributions have a long tail that has nothing to do with incompetence on the user's side. As a rule of thumb, a quarter of users will need about twice the median time to solve the task.

Acknowledgments

I would like to thank Jim Lewis and Joe Dumas for their encouragement to keep going with this research, and my SAP colleagues for thousands of hours of lab time collecting the data.

References

Albert, W., Tullis, T., & Tedesco, D. (2010). *Beyond the usability lab: Conducting large-scale online user experience studies.* Burlington, MA: Morgan Kaufmann Publishers.

Coursaris, C., & Kim, D. (2011). A meta-analytical review of empirical mobile usability studies. *Journal of Usability Studies, 6*(3), 117–171.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, *17*(1), 111–117.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79–102.

ISO (2006). *Software engineering–Software product quality requirements and evaluation (SQuaRE)–Common Industry Format (CIF) for usability test reports.* ISO/IEC 25062:2006(E).

ISO (2013). *Usability of consumer products and products for public use — Part 2: Summative test method.* ISO/TS 20282-2:2013(E)

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data.* New York: Springer

Liu, C., White, R. W., & Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 379-386). New York: ACM.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford Psychology Series No.8.

Molich, R., Chattratichart, J., Hinkle, V., Jensen, J. J., Kirakowski, J., Sauro, J., … Traynor, B. (2010). Rent a car in just 0, 60, 240 or 1,217 seconds? – Comparative usability measurement, CUE-8. *Journal of Usability Studies, 6*(1), 8–24.

NIST/SEMATECH (2012a).Empirical model fitting—Distribution free (Kaplan-Meier) approach. In E-handbook of statistical methods. *National Institute of Standards and Technology*. Retrieved December 2013, from http://www.itl.nist.gov/div898/handbook/apr/section2/apr215.htm#Modified K – M.

NIST/SEMATECH (2012b).Critical values of the normal PPCC distribution. In E-handbook of statistical methods. *National Institute of Standards and Technology*. Retrieved December 2013, from http://www.itl.nist.gov/div898/handbook/eda/section3/eda3676.htm.

Rummel, B. (2014). Probability plotting: A tool for analyzing task completion times. *Journal of Usability Studies 9*(4), 152–172*.* Retrieved from http://uxpajournal.org/probability-plotting-a-tool-for-analyzing-task-completion-times-2/

Sauro, J. (2011). 10 things to know about task times*. Measuring Usability*. Retrieved December 2013, from http://www.measuringusability.com/blog/task-times.php

Sauro, J., & Lewis, J. R. (2009). Correlations among prototypical usability metrics: Evidence for the construct of usability. *CHI '09 Proceedings of the SIGCHI Conference on Human Factors in Computing System* (pp. 1609–1618). New York, NY: ACM Press.

Sauro, J., & Lewis, J. R. (2010). Average task times in usability tests: What to report? *CHI '10 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2347–2350). New York, NY: ACM Press.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience.* Waltham, MA: Morgan Kaufmann.

Tobias, P. A., & Trindade, D. C. (2012). *Applied reliability* (3rd ed.). Boca Raton, FL: CRC Press.

About the Author

Bernard Rummel
Mr. Rummel is trained in experimental psychology and has been working in the Usability and UI Design field for over 20 years. After nine years at the German Naval Medical Institute, he joined SAP in 2000, where he is currently responsible for usability testing methodology and participating in the national standardization body DIN.

Appendix

I have provided a specialized spreadsheet calculator so readers can perform Weibull analyses on their own for up to 2000 task times; it can be extended for larger datasets. The template opens separate workbooks that require Excel 2010 or a later version. For automatically optimizing t_0 estimates, the Excel *Solver* Add-In is required. To load Solver, select File > Options > Add-Ins > Manage > Solver Add-in.

The workbook implements modified K-M estimates for task completion rates < 1 and also contains critical R^2 values, as provided by NIST/SEMATECH (2012a and b). The workbook contains two separate spreadsheets:

- The first sheet, *Analysis*, contains calculation and charting templates for probability plots to analyze exponential and Weibull distributions. Users can paste in their own data and follow further instructions in the sheet. For the Weibull distribution, the workbook will calculate parameters from the data entered and evaluate the model fit. Users can also enter target times or task completion rates, the spreadsheet then will calculate corresponding expected completion rate or time estimates.
- The second sheet, *Significance Levels Table*, contains a significance table listing critical R^2 values for different sample sizes. This table should not be removed or changed; it is used for looking up the significance assessment of the model fit.

For analyzing normal and lognormal distributions, and more general analysis scenarios, refer to Rummel (2014) where those are discussed and a more generic spreadsheet is provided.