ASSESSMENT OF INTERRATER RELIABILITY AMONG *C. ELEGANS* RESEARCHERS MEASURING DEVELOPMENTAL STAGE BY THE KAPPA STATISTIC AND LATENT VARIABLE MODELING

by

Annabel Ansel Ferguson

BS, BA, University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Annabel Ansel Ferguson

It was defended on

November 27th, 2013

and approved by

Thesis Advisor: Gary M. Marsh, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Richard A. Bilonick, PhD, Assistant Professor of Ophthalmology, Biostatistician, School of Medicine, University of Pittsburgh

Committee Member: Jeanine Buchanich, PhD, Research Assistant Professor, Graduate School of Public Health, University of Pittsburgh

Committee Member: Alfred Fisher, MD, PhD, Associate Professor of Medicine, University of Texas San Antonio

Gary Marsh, PhD

ASSESSMENT OF INTERRATER RELIABILITY AMONG *C. ELEGANS* RESEARCHERS MEASURING DEVELOPMENTAL STAGE BY THE KAPPA STATISTIC AND LATENT VARIABLE MODELING

Annabel Ansel Ferguson, MS

University of Pittsburgh, 2013

ABSTRACT

Use of the tiny nematode worm *Caenorhabditis elegans* as a model organism for biological research has had a considerable influence on scientific discoveries. *C. elegans* research has a public health relevance as it has led to better understanding and treatments of diseases like cancer and neurodegenerative diseases, which are a public health concern. Additionally, *C. elegans* research offers promise towards a better understanding of the public health problems of human obesity and diabetes, as many initial controlled experiments may only be done in a model organism (as opposed to human studies). A commonly employed skill among *C. elegans* researchers is the ability to reliably and accurately distinguish among the five stages of worm development by eye; this is required for both producing quality data, and for successful worm maintenance and genomic manipulation. While it is reasonable to presume that there is some amount of variability from researcher to researcher in classifying worms into particular stages, there is little documentation of formal assessments of reliability between researchers.

The topic of statistical assessment of interrater reliability has been addressed extensively as it applies to fields like medical diagnostics, and psychological and sociological studies. While numerous methods exist, a popular way of assessing the reliability of two or more different measurements on a categorical scale is the kappa statistic. The ease of computation and the single numerical index (ranging from 0 to 1) of the kappa make it a commonly used “quick” method of this assessment. However, there are numerous problems that arise in the interpretation and application of kappa that can make it untrustworthy, especially when it is used for the analysis of data with an ordinal outcome such as developmental stage.

An alternative methodology is latent variable modeling using a common factor model with a probit transformation. This approach provides more information about the strength of association and bias among raters, and does not give the potentially confusing or paradoxical results that kappa does. The following study has applied both of these methods to a dataset containing the ratings of worm larval stage of development for a population of 60 worms by seven raters. This study finds that while both the kappa and the modeling approach give concordant results, modeling the data provides a more useful and meaningful summary of the agreement between raters. Additionally, it finds that the overall agreement is high, but that there is some degree of variability in the cutoff thresholds by which raters assign developmental stage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

PREFACE

This work was only possible with the expertise and resources provided by the thesis committee members Dr. Gary Marsh, Dr. Richard A. Bilonick, Dr. Jeanine Buchanich, and Dr. Alfred Fisher. Dr. Fisher provided the lab space and equipment necessary for gathering the raw data. The post docs, graduate students, and technicians in the Fisher lab who kindly and willingly participated in this study played a crucial role in this study; without their participation and *C. elegans* experience, the dataset could not exist. In addition, Dr. Bilonick provided the code and expertise necessary for running the structural equations modeling.

1.0 INTRODUCTION

In most types of research, when one wishes to draw a general conclusion about a particular population or the difference between two populations, it is common practice to make measurements on a subset of the population. Summary statistics like mean and standard deviation are used to make general statements about the magnitude and variability of the particular phenomenon being measured in the population. When investigators make conclusions about significance there is an implicit assumption that the magnitude and variability within the dataset are due to the natural magnitude and variability present in the population being measured. However, it is well known that these statistics may be distorted by systematic errors (biases) and random errors that arise from the method of measurement itself. In certain types of experimentation or medical diagnostics, the degree to which measurement error or bias has an impact on overall statistics is of interest. This would especially be the case when adequate experimental control for measurement error is not possible; for example, if all measurements were not being made by the same technician. The degree of measurement error present in a particular test may impact the conclusions made from a study. For example, in a simple regression problem where the expected value of the response Y is a linear function of a predictor X, say $\alpha + \beta X$, it is assumed that X is known. It is well known that when X must be measured and is subject to random errors, the estimate of β will be biased toward zero and the larger the random errors, the larger the bias (1). When there is concern that measurement error is present, one may conduct repeated measurements either using the same instrument, or using different instruments so that the measurement error can be separated from true variability. Often times it is of interest to compare instruments or methods solely for the purpose of assessing the nature of systematic differences and any differences in the expected magnitudes of the random errors exhibited by each instrument. Further, the “instruments” may consist of experts making judgments or ratings of some attribute, and these ratings are measurements on a discrete ordinal scale. The goal may be to determine how best to improve agreement among the raters. Assessing interrater agreement in such cases is the topic of interest in what follows.

This project will provide a background on two different approaches to assessing interrater reliability: the first approach is the commonly used kappa statistic, and the second approach is through latent variable modeling. In addition, interrater assessment will be made using these two approaches for a particular problem in biological research done using the model organism *Caenorhabditis elegans*.

1.1 INTERRATER RELIABILITY

A major challenge for research in the natural sciences, psychology, sociology, and medical diagnostics is in assessing how reliably two or more measurement techniques evaluate the same phenomenon. This is particularly vexing in cases where a well-defined gold standard measurement does not exist for the outcome of interest, and one cannot confidently say that a particular measurement technique is more accurate than the others. For instance, in medicine, physicians will typically diagnose a disease in a patient based on the presentation of a set of symptoms; the ultimate decision of whether or not a patient has a disease is a judgment made by the physician, and in turn, these will vary from physician to physician. A certain degree of consistency is desired between doctors to ensure 1) that diseases are not being over- or under-diagnosed in clinics, and patients are receiving adequate and appropriate care, 2) that any research studies addressing the disease of interest possesses internal validity, and 3) that estimations for disease prevalence and incidence are accurately reported for public health monitoring. An example of measurements in which reliability between raters, or interrater reliability, is a particular concern include clinical psychological diagnosis of patients into categories such as “psychotic”, “neurotic” or “organic” between physicians (2). Concerns over interrater reliability may be present for categorical data with two or more outcomes, or for continuous data. They may also be present over comparisons between rating systems with differing numbers of categories. However, this project will focus on ordinal categorical ratings, with an equal number of outcomes between raters. It will address how to measure the reliability between two or more raters, with a rating system consisting of two or more ordered categories.

1.1.1 Kappa Statistic

A comparison between two raters for an outcome with two categories may be displayed in a 2x2 table as shown in Table 1.

Table 1. 2X2 table for two raters

		Rater 1 (x_1)		
		yes	no	
Rater 2 (x_2)	yes	a	b	a+b
	no	c	d	c+d
		a+c	b+d	N

The variables a, and d represent the counts for which the two raters agreed on the outcome, and the values b, and c are discordant between these two raters. As expected, tables with higher values for a, and d and lower values for b, and c have a greater degree of interrater reliability. A simple way to represent this agreement is by stating the proportion of the total that are in agreement, or $P_o = (a + d)/N$ (2). Another commonly used value to indicate degree of similarity is the intraclass correlation, ρ, which is shown in Equation 1. The highest value of ρ is one, which indicates perfect agreement and occurs when the discordant cells (c and b) are zero. The lowest value of ρ is minus one, which indicates perfect disagreement, which occurs when a and d are zero.

$$\rho = \frac{4ad - (b+c)^2}{(2a+b+c)(2d+b+c)}$$

Equation 1.

A limitation to using either of these values to represent interrater reliability is that they do not account for the fact that two raters may have some amount of agreement by random chance. The kappa statistic was introduced to address this issue, and present a measure of agreement that has been chance-corrected, by removing the proportion of agreement expected in the 2x2 table by random chance. Equation 2 represents the simplest form of the Kappa statistic, or Cohen's Kappa (3).

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Equation 2.

$$P_o = \frac{a + d}{N}$$

Equation 3.

$$P_e = \frac{(a+b)(a+c) + (c+d)(b+d)}{N^2}$$

Equation 4.

The ease of computation, and simple interpretation of kappa has made it a widely used method for reporting agreement between raters (3, 4). Specifically, because kappa is chance-corrected, a value of one may be interpreted as "perfect" agreement, and a value of 0 means that the agreement between the raters is no better than agreement by chance alone; any value greater than 0 signifies some degree of agreement between raters (4). As a guide for interpretation of kappa values in terms of the strength of agreement, Landis, et. al. have proposed the following guidelines: kappa ≤0 is poor agreement, 0.01 – 0.2 is slight, 0.21 – 0.4 is fair, 0.41 – 0.6 is moderate, 0.61 – 0.8 is substantial, and 0.81 – 1.0 is almost perfect agreement (5).

The kappa statistic may be extended to measure agreement between multiple different raters assessing an outcome with more than two categories. To extend kappa to multiple rating categories (shown in the table in the k x k table in Table 2), the same expression for kappa in equation 2 is used, except the expressions in Equations 5 and 6 are used for P_o and P_e, respectively.

$$P_o = \sum_{i=1}^{k} p_{ii}$$

Equation 5.

$$P_e = \sum_{i=1}^{k} p_{*i} p_{i*}$$

Equation 6.

Table 2. kXk table for two raters

		Rater 1 (x_1)				
		1	2	...	k	
Rater 2 (x_2)	1	p_{11}	p_{12}	...	p_{1k}	p_{1*}
	2	p_{21}	p_{22}	...	p_{2k}	p_{2*}
	...	...	...	...	...	...
	k	p_{k1}	p_{k2}	...	p_{kk}	p_{k*}
	total	p_{*1}	p_{*2}	...	p_{*k}	N

When the number of ratings is constant, an overall kappa value may be obtained, as described by Landis et. al., and Fleiss, et. al. (2, 5).

For each individual category j, a kappa value may be calculated with the following expression in equation 7 and equation 8, where n = number of subjects, k = number of categories, x_{ij} are the observed number of ratings for each subject (i) and category (j), m is the sum of the ratings over all categories, p_j = overall proportion of ratings in category j and q_j = 1 - p_j.

$$\kappa_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)p_j q_j}$$

Equation 7.

$$m = \sum_{j=1}^{k} x_{ij}$$

Equation 8.

$$\kappa_{overall} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} p_j q_j}$$

Equation 9.

Kappa was originally intended for analyzing nominal categorical data, without any notion of one category being more similar to another. With ordinal data however, due to the ordered nature of the categories one would expect that some discordant ratings are more severe than others. For example, if there are four ordered categories, 1, 2, 3, and 4, then a disagreement between categories adjacent to each other (1 and 2, 3 and 4, for instance) is a lesser disagreement than when raters are discordant between categories that are not adjacent (1 and 3, or 1 and 4). Cohen proposed the weighted kappa statistic as a way to reflect this information (4, 6). Relative weights are attached to cells of a k by k table (Table 2), with a value of one given to the concordant cells, and a value between zero and one given to the discordant cells; the weights given to the cells are chosen arbitrarily.

1.1.2 Shortcomings of Kappa

Since its introduction, the use of kappa has also been extensively criticized. The main problems include issues with interpretability and generalizability of kappa(7-10). Specifically, that 1) there is a lack of consensus about what value of kappa is high enough to be considered good agreement among raters, 2) the value of kappa is dependent on the prevalence of the trait being measured in a population, and may take on paradoxically high or low values 3) that kappa is an omnibus index comprised of many components of rater disagreement, and 4) that when there is an ordering to the categories, the assignment of weights to the categories is arbitrary, and may not reflect the true relative differences between categories.

The first issue arises from the interpretation of a kappa value that lies between zero and one. Significance is usually determined to test the null hypothesis of kappa being equal to zero (testing whether there is agreement that is not due to chance alone), however, investigators typically want to know whether their raters have a sufficiently high level of agreement, and merely having agreement that is better than chance is not good enough. While it is possible to designate cutoff values for kappa values corresponding to high agreement, a decision about where to draw these cutoffs would have to be made on an arbitrary basis (10). It is also conceivable for kappa to take on a value of less than zero, which is difficult to interpret. The second issue arises from the notion that kappa may vary with the prevalence of the underlying trait; this issue is also known as the base rate problem (7, 9-11). This is illustrated by Byrt et. al. using two by two tables that have the same observed proportion of agreement, but the calculated kappa values differ due to the distribution in the marginal totals. Tables 3 and 4 recapitulate this illustration.

Table 3. Comparable kappa and proportion of agreement

A\B	Y	N	total
Y	40	9	49
N	6	45	51
	46	54	100

$P_o = 0.85; \kappa = 0.70$

Table 4. Small kappa and high proportion of agreement

A\B	Y	N	total
Y	80	10	90
N	5	5	10
	85	15	100

$P_o = 0.85; \kappa = 0.32$

Tables 3 and 4 both have the same P_o's, of 0.85, however the kappa values are 0.70, and 0.32 for Table 3 and Table 4, respectively. This demonstrates an instance in which kappa may vary considerably, depending on the distribution of the data. Table 3 has a fairly even distribution of yes and no answers in the marginal totals, while Table 4 has a much higher proportion of yes answers in the marginal totals, which likely reflects that there is a higher prevalence of the disease within the population being measured. The difference between the kappa values for these two tables is due to a difference in the expected proportion of agreement, which is 0.5 for Table 3 and 0.78 for Table 4. This example raises the notion that in some cases, the comparison of kappa values obtained from raters assessing different pools of subjects may not be reliable. The fact that the kappa value in Table 4 of 0.32 suggests low agreement even though the proportion of agreement is very high (0.85) is worrisome because it contradicts the initial first glance results. This has been called a paradoxical result of using kappa (7-9).

While this example is troublesome, this sort of paradoxical result only becomes an issue when the prevalence of the trait being measured in a population is extremely high or low. A detailed analysis of the behavior of kappa under varying rater sensitivities, specificities, and population trait prevalence was described by Guggenmoos-Holzmann (9, 10, 12). This author demonstrates that kappa is extremely unreliable as the trait prevalence approaches zero or 1. This study shows that kappa becomes more stable for populations with a trait prevalence between 0.3 and 0.7, but that this stability also depends on the sensitivities and specificities of the different raters, which indicates that there is always some amount of uncertainty in how reliable kappa is, and to what degree kappa is actually influenced by the trait prevalence.

The third issue of kappa is related to interpretability of the kappa value, in light of the fact that kappa is a composite value, representing several different aspects of rater agreement. This issue is related to issue 2, in that trait prevalence is one component on which kappa depends. In addition to the prevalence and variance of the trait being measured, kappa is also influenced by rater bias and rater error (7-10, 12). In assessing rater reliability for categorical measurements with more than two categories, one rater may have a tendency to classify items in one category more than the other raters, which is an example of rater bias. On the other hand, one rater may differ from another due to random error, without any directionality to the disagreement. The kappa statistic alone does not make a distinction between these two sources of disagreement. This reduction of information in the kappa value makes it more difficult for an investigator to make specific suggestions for how the interrater reliability may be improved in a particular study.

The fourth issue is present for data in which the rating categories have a natural ordering. To use kappa to describe agreement in such data, one may choose to use the basic form of kappa that does not account for certain discordances being larger than others, and this information is not represented by the statistic. Alternatively, one may assign weights to the categories, and calculate a weighted kappa based on these assignments. This approach is problematic because the weights are chosen arbitrarily and this form of kappa may be misleading because it assumes that the relative distances from one rating category to another is known (10).

Many investigators have made the case for why reporting kappa by itself, without additional summary statistics or raw values for the individual raters, is an insufficient representation of interrater reliability. Some authors have recommended that kappa be reported along with other indices that adjust kappa for underlying prevalence and balance of the data (7, 10, 12, 13). Other authors have stated that kappa is not very useful, and that other methods, like latent variable modeling ought to be used instead to analyze agreement (10, 12, 14).

1.2 LATENT VARIABLE MODELING WITH STRUCTURAL EQUATIONS

Latent variable modeling offers a more robust and consistent means of assessing interrater reliability, however this approach is also much more computationally demanding and conceptually challenging to understand. The general utility of a latent variable modeling approach is that it addresses situations where a phenomenon or physical quality cannot be directly measured, and a representative measurement is made instead. This approach is a means to infer specific properties of un-observable variables from variables that are measureable (15). The un-observable variables are also known as latent variables, while the measurable variables are called manifest variables. An example of a latent variable that would be of interest to an investigator might be disease severity of depression, which presumably follows a continuum, but is also not directly measurable. A manifest variable that represents depression severity and also is directly measurable, would be some value along a depression severity rating scale, such as the

Hamilton Rating Scale for Depression, obtained from an individual's response to a questionnaire (16). In this case, the ordinal value obtained from the questionnaire ought to be correlated to the "true" latent variable for depression severity. If an investigator were interested in assessing the relationships between depression and other factors (such as income, age, family history, etc.), one could build a model linking the measured and unmeasured variables, and ultimately obtain estimates for the parameters in the model. In order to build a good model, there are several key concepts one must understand that will be introduced in the sections below. These include: path diagrams, structural equations models, and identification of a model. Additionally, for the assessment of ordered categorical data, the method of ordinal regression will be introduced.

1.2.1 Path Diagrams

The geneticist Sewall Wright invented path analysis in 1921 (17). A path diagram refers to a specific graphical display that represents interrelationships between many variables(17). These diagrams are meant to provide a visual display of potentially complicated relationships between different factors, which can make conceptualization of a particular system feasible. By using meaningful symbols and notations, a path diagram will translate into an exact set of equations whose parameters may be estimated(17, 18). While a path diagram itself is not essential to the analytics and acquisition of model parameters, it is useful for conceptualization and comprehension of what each variable and parameter refers to in a model. Figure 1 is an example of a path diagram.

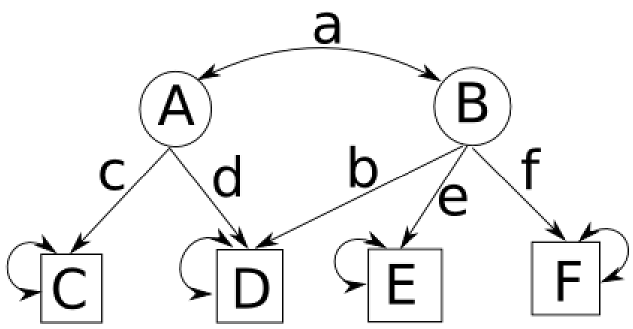

Figure 1. Path diagram with latent and manifest variables

In Figure 1, there are six variables, represented by uppercase letters A, B, C, D, E and F. The manifest, or observed variables are outlined in boxes (C, D, E, and F), while the latent variables are circled (A and B). The arrows connecting the variables are either curved with two heads, or straight with one head. The curved two-headed arrow designates a correlation between two variables (18). The straight, one-headed arrow connecting two different variables signifies a linearly dependent relationship between the two variables, with the arrowhead pointing in the direction in the dependency. This means that a change in the variable at the tail of the arrow will result in a change in the variable at the head of the arrow, but not vice versa (18). The lowercase letters are the magnitudes of the strength of the correlation between two variables, or the slope of the linear relation. Finally, the curved two-headed arrow linking a variable to itself is used to represent an error or disturbance term, which may account for errors within a variable that are not accounted for by the other terms in the model. In Figure 1, latent variables A and B are correlated with each other, and manifest variables C and D are both dependent on A; the manifest variables all have error terms.

1.2.2 Structural Equations.

One way of expressing the information contained in path diagrams is through a set of structural equations. These are constructed by equating each downstream variable to a function of the paths leading to it, such that each equation contains a term for each straight arrow leading to the downstream variable (18). These equations are essentially regression equations, where the lowercase letters are regression coefficients. For instance, structural equations representing the path diagram from Figure 1 are as follows:

$$C = cA$$
$$D = dA + bB$$
$$E = eB$$
$$F = fB$$

In order to make a comparison between regression coefficients for different equations in the path diagram, the coefficients may be standardized by multiplying the rawscore coefficient by a ratio of the standard deviations of the variable at the tail of the arrow to that at the head of the arrow(18).

A critical aspect of structural equations modeling is the ability to estimate model parameters using an observed dataset. This is dependent on the number of structural equations, and the number of known and unknown variables. If there are too few equations and known parameters to estimate the unknowns, then the model is underdetermined. A model that has the minimum number of unknown parameters in order to provide a unique solution is just determined, or identified (18)p. 16. To show that model identification has been reached, it is necessary to demonstrate that each unknown parameter may be expressed as a function of known parameters with a unique solution (17)p. 88. Over determination, or over identification of a model refers to cases where there is an excess of information with which to estimate model parameters. If at least one of the unknown parameters can be expressed with more than one function of known parameters, then the model is over identified (17)p. 90. An over-identified model is usually preferred since each measured variable has its own measurement error that has a bias from the true value. A redundancy of identifying data for the unknown factors will usually give a closer estimate of the true underlying value (18)p. 17.

1.2.3 Factor Analysis.

Path diagrams provide a useful means of estimating parameters found in factor analysis. Factor analysis, as originally devised by Spearman in 1904, refers to the idea of several measured variables being related to each other by sharing causal relations to one common factor. Typically, factor analysis is used to relate a number of measured factors to a single unmeasured, or latent factor. An example path diagram for factor analysis is shown in Figure 2.

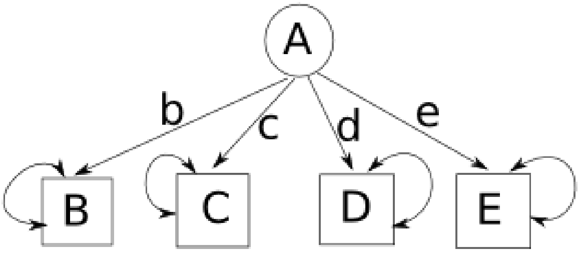

Figure 2. Path diagram for a one-factor analysis

Factor analysis may be used in an exploratory, hypothesis generating manner, in which different models are tested, and variables are added or eliminated, depending on model fit. This form of factor analysis is known as exploratory factor analysis. In confirmatory factor analysis, on the other hand, the model has already been proposed based on an existing theory, and the number of factors or variables, and their relationship to each other is already stated. The main aim of confirmatory factor analysis is to determine how well the interrelations among the measured variables are accounted for by the model (19).

In order to model interrater reliability, a confirmatory factor analysis model may be used. The path diagram in Figure 2, for example, could represent a model for a comparison between four different raters. In this model, the four raters represented by variables B, C, D, and E are assessing the same phenomenon, which is represented by the latent variable A. The measurements made by each rater are independent of each other, but have a linear dependency on the same underlying factor.

Uebersax describes a latent trait agreement analysis approach, which may be used to model interrater agreement for categorical ratings(14). This approach assumes that the latent trait underlying the ratings is continuous, and that each rater has specific thresholds along the continuum to determine how he or she categorizes an item (14). The parameter estimates for this type of model give quantities that represent rater bias, category definitions, as well as measurement error. These parameter estimates each describe how closely each rater is able to measure the common latent factor.

For the comparison of raters using an ordinal rating scale with more than two categories, a factor analysis model is appropriate, given that each rater is assessing the same latent quality. Such a model could have a path diagram similar to that in Figure 2, with B, C, D, and E each being a different rater, and A being the latent quality each rater is assessing. However, if one assumes that the latent quality follows a continuous distribution, than additional steps will be needed in order to relate the ordinal outcomes given by the raters to the latent variable, and estimate parameters to explain the relationships between raters. Specifically, one must determine what sort of distribution the latent trait follows, and also, the ordinal outcome data from the raters must be transformed such that it bears a linear relation to the latent trait. The topic of Ordinal Regression covers methods of relating an ordinal categorical outcome variable to a continuous predictor. When attempting to analyze data that has an ordinal scale by classical regression analysis, one can either treat each ordinal category as a separate indicator variable, or one may treat the ordinal data as continuous, using the numerical values assigned to each category in the model. Both of these approaches are problematic. In treating each ordinal category as an indicator variable, one ignores the ordered nature of the categories, and there is a loss of information in this analysis. By treating the ordinal data as a continuous scale, one incorrectly assumes that the relative distances between each category are determined by the numerical values assigned to each category, which can give misleading conclusions (20).

In ordinal regression one assumes that the ordered categories represent different sections along a continuous distribution that are delineated by threshold values. For example, Figure 3 shows a normal distribution with two threshold values x_1 and x_2. The letters A, B, and C are the ordinal categories into which the data are classified. Any latent value that is lower than threshold x_1 will fall in category A, while a latent value lying in between x_1 and x_2 will be category B, and anything greater than x_2 is C.

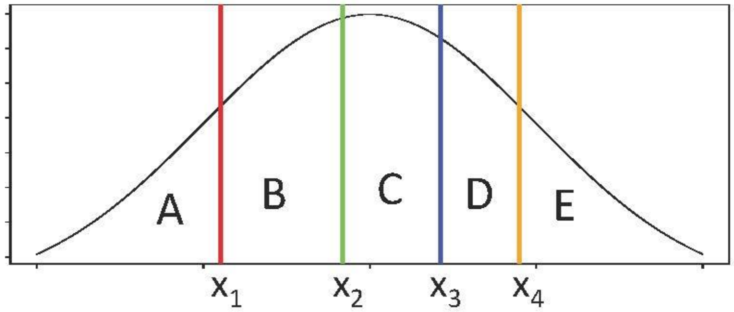

Figure 3. Standard normal distribution with thresholds.

To appropriately fit a one factor model to data that has ordinal outcomes, the data must be transformed such that the category values represent different levels along a normal distribution. The link function that transforms the data is called a probit function, and it is introduced by Agresti, and others (20, 21). A path diagram for a one factor model with ordinal data is shown in Figure 4. In this model there are three raters who share a common underlying factor. Each rater has their own latent distribution that determines the ordinal ratings, and this distribution is connected to the ordinal values by the squiggly arrow, which represents the probit transformation. The threshold values designating each level may be freely estimated, and may vary from rater to rater due to the fact that each rater may have slightly different ways of defining each category.

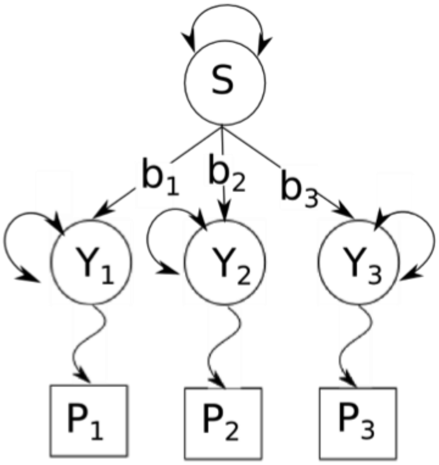

Figure 4. Path diagram for a one-factor ordinal model

In this model, the parameters of interest are the threshold values for category classifications between the raters (the number of threshold parameter estimates is equal to the number of categories minus one, for individual each rater), and also b_1, b_2, and b_3, which describe the magnitude of the linear association between each rater's distribution, and the latent distribution from the population that each rater is measuring. In this model, the underlying distributions for the latent terms have been constrained to following the standard normal distribution. Therefore, with this standardization, the b terms, which are also called factor loadings, are correlation coefficients between each rater and the common factor. Furthermore, the amount by which each rater's factor loading is smaller than one represents the amount of random error for this rater. From this model, one may compare a set of raters, and determine whether one or more rater has a low b estimate, meaning that this rater has a relatively high amount of random error. The threshold values, may also be compared between raters, to see whether one rater has different category definitions from the other raters. For this project, a model similar to Figure 3 will be used, except it will be scaled up to 7 raters, and an outcome with 5 categories.

1.3 *CAENORHABDITIS ELEGANS* BACKGROUND

The tiny transparent non-parasitic nematode worm, *Caenorhabditis elegans* has been a model organism for biological studies since the 1960's. Presently, there are approximately 21,500 *C. elegans* researchers listed on wormbase.org, conducting experiments in topics including genetics, aging, pharmacology, neurobiology, and ecology. A number of qualities make worms a useful system in which to study biological processes. Stocks are easy and inexpensive to maintain since worms have a three-day lifecycle and a one-month lifespan, they reproduce sexually as self-fertilizing hermaphrodites meaning that a worm strain may be revived with a single fertile hermaphrodite worm, and adult worms are only one millimeter in length. Additionally, the large worm research community that has existed for over 50 years has afforded a wealth of data and resources, including a fully sequenced and extensively mapped genome, a large library of genetic mutant strains, an RNAi library enabling silencing of almost any known gene, and an online database containing information on the worm genome, phenotypes, reagent information, bibliographies, worm morphology, etc. (22).

The *C. elegans* life-cycle consists of development as an embryo, then following hatching, the progression through the L1, L2, L3 and L4 larval stage, and a final molt into the reproductive adult stage. Figure 5 shows the lifecycle of *C. elegans*. Under normal, optimal growth conditions, a worm will undergo the normal development pattern (embryo, L1, L2, L3, L4, and adult) in 3 days. The worm also has the capability of undergoing an alternative life-cycle pattern when confronted with stressful environmental conditions, including overcrowding, a lack of food, and high temperature. Such conditions may trigger the diapause life-cycle, in which a worm skips the normal L2 and L3 stage of development, and instead becomes a L2d (predauer) followed by dauer, and then L4 and adult. The dauer stage of development is characterized by physiological and morphological changes that enable the worm to survive a harsh environment, including an increased store of fat, an impervious and stronger cuticle, the ability to move rapidly, and a decrease in metabolism (23).

normal: L1 → L2 → L3 → L4 → reproductive adult
diapause: L1 → L2d → dauer → L4 → reproductive adult

Figure 5. *C. elegans* life cycle

A considerable amount of *C. elegans* research has involved study of the dauer larvae. Experiments that have screened for mutations in the worm causing an inappropriate entry into dauer have yielded stress response genes, and genes that additionally affect worm lifespan, including the worm insulin-like signaling receptor *daf-2*. The overlap between genes that influence dauer entry and genes that influence longevity, as well as the conservation of the signaling pathways that determine dauer decision, make study of the dauer larvae an important aspect of *C. elegans* research.

The correct identification of a dauer larva, and the ability to distinguish between the different stages of worm development are necessary skills for a worm researcher to have; both for the routine maintenance of stocks, and for the generation of accurate and reliable data. A number of publications have reported the proportion of dauer larvae in a population as a read-out for the inappropriate dauer entry phenotype (22, 24-28). These studies have identified important genes relevant to aging, diabetes, and cancer. Included in the pool of C. elegans genes that are involved in dauer development are *daf-2* (the insulin/IGF-1 receptor) (28, 29), *daf-16* (the FOXO transcription factor) (30), as well as *akt-1*, *akt-2*, and *age-1* (genes involved in insulin signaling) (31, 32). Some of these studies depend on the experimenter scoring the worm population by eye (33). It is expected that such experiments are subject to bias from one rater to another, however, there is little documentation for formal analyses of rater reliability in the worm community. While other, less biased, methods exist for assessing dauer proportion (such as SDS selection, which consists of rinsing the worms with a solution that kills all stages of development except for dauer), scoring a population by eye allows for the assessment of other stages of development resulting in higher resolution. In addition, treatment with SDS will also kill both partial dauer stages and pre-dauer stages of development; these stages are sometimes highly prevalent in a population of worms, and if a researcher wants to measure these types of dauers, it is necessary to score by eye (34). A means of formally assessing reliability and bias among worm researchers could provide a standardization for developmental assay experiments, as well as aid in the training of less experienced researchers. If a worm lab can demonstrate that each lab member has the same propensity for rating a particular stage of development, this would provide evidence for internal validity of development assays scored by eye.

2.0 METHODS

2.1 DATA COLLECTION

To obtain a mixed stage population of worms containing dauer larvae as well as L1, L2, L3 and L4 stages, wild type N2 worms were grown at 20°C on nematode growth agar (NGA); this population was supplemented with dauer larvae that were transferred from an *eak-4;tatn-1* double mutant stock, grown at 25°C. The *eak-4;tatn-1* double mutant will consist of approximately 90% dauer larva when grown at 25°C. Movies of the mix stage NGA plate were obtained using an I.C. capture 2.0 microscope, with a video camera attached. These moving images were taken at a 40X magnification. There were 9 separate 5 to 30 second movie files made; each movie containing 3 to 10 worms that a rater would score. To ensure that all raters were scoring the same individual worms, the worms from these movies that were rated were each given a unique identifying number; these were labeled in a separate image of a still from the first frame of each movie. An example is shown in Figure 6 below.

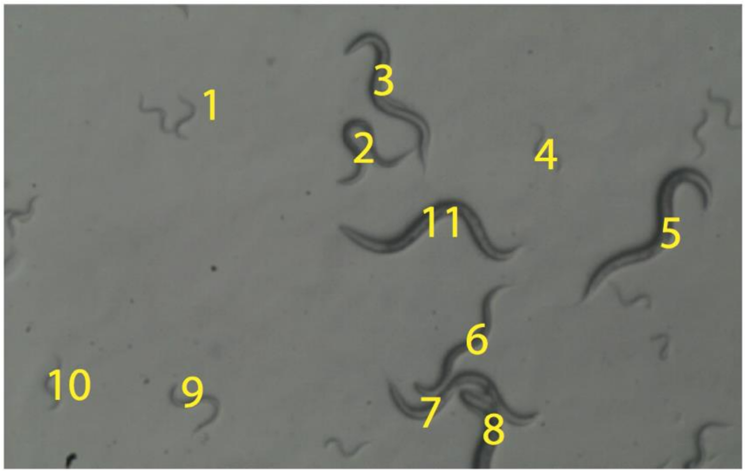

Figure 6. First frame of a movie containing worms numbered 1 through 10

Worm rater data was obtained by recruiting seven individuals who had some amount of C. elegans research experience, which ranged from less than a year to nine years of experience. Before rating the worms, each rater was given a five-minute tutorial by the author, explaining the qualifying features of each stage of development, and showing an example video containing each stage. Raters then were given a score-card (Appendix A), and for each movie file, were asked to identify the particular worm to be rated, and then play the video as needed in order to make a decision on the stage of development, making a mark under the stage of development column corresponding to each worm. This decision was typically reached after only a second or two of viewing the worm; indeed, a rating was usually given after only viewing the still frame from the movie.

These data were tabulated with a column for each rater. A total of seven raters were recruited for this study, and each rater was asked to judge the larval stage of 50 different worms. Hence each rater had 60 observations, with the larval stage of development coded by ordinal values one through five. Each rater had some amount of *C. elegans* experience, ranging from nine years to less than a year. Rater ALF had the most expertise, with over nine years of experience, followed by AAF with three to six years of experience, SAK, AGS, and DAH, who each had one to three years of experience, and raters UMA and HNW each had less than a year of experience. There was one missing value in this dataset, which was observation number 41 by rater DAH.

Raters were coded by initials as follows:

AAF = 1
ALF = 2
AGS = 3
DAH = 4
HNW = 5
SAK = 6
UMA = 7

The larval stages were coded as follows:

L1 = 1
L2 = 2
dauer = 3
L3 = 4
L4 = 5

2.2 ANALYSIS OF INTERRATER RELIABILITY

2.2.1 Plots and descriptive statistics

The pairwise error plots were graphed by plotting the results of one rater against the results of another, and displayed in an array. These were produced using the merror.pairs statement in R software, and a small amount of noise was added to the data with the jitter statement so that relative point densities could be compared across each graph. The observed frequencies and histograms were produced using STATA.

2.2.2 Kappa Calculations

Kappa statistics were calculated using the kap command in STATA software, which calculates the kappa value when each rater's outcomes are listed in a column. For the pairwise kappa values, each pair was coded as follows: `kap raterA raterB`. The resulting values for observed and expected probabilities (p_o and p_e) correspond to those values in Equations 5 and 6, and the pairwise kappa value was calculated according to Equation 2 (3). The command for getting the multi-rater kappa for all seven raters was stated as follows: `kap rater1 rater2 rater3 rater4 rater5 rater6 rater7`. This multi-rater kappa value was calculated as described by Fleiss et. al., according to Equations 7 8 and 9, where the kappa values comparing raters' agreement for single categories were obtained from Equation 7, and the overall kappa was obtained from Equation 9 (2).

2.2.3 Latent Variable Modeling

Latent variable modeling was conducted with R software, OpenMX package. The one factor ordinal model with seven raters and five ordinal categories was adapted from code available from (http://openmx.psyc.virginia.edu/docs/OpenMx/latest/FactorAnalysisOrdinal_Matrix.html). An annotated version of the code used to produce the final results is shown in Appendix C. The parameters were estimated using a -2 log likelihood process. Due to the complexity of this model (having seven raters and five categories), and possibly because there were only 60 observations, there was a certain degree of computational challenge in achieving a minimum -2 log likelihood. The parameter estimates given by openMx came with a “code RED” warning message that stated the minimum -2LL may not have been achieved. In order to gain more confidence that the best parameter estimates were reached for this model, it was necessary to run the code using different starting values for the parameter estimates. This process was repeated over 100 times using slightly altered starting values for each run, and the parameter estimates from the model giving the lowest -2LL value were used in the final report. Appendix C shows 10 attempts at running this model that had the lowest -2LL, with starting values, the -2 log likelihoods and the parameter estimates. For this report, the results shown were generated from manually entering different starting values, and running the program many times. The best results from these attempts had a -2LL of 825.841. However, a code was also generated that automated the process of repeating the model with different starting values for each run, as shown in Appendix D. In this code, the starting values for the parameter estimates were chosen at random from a uniform distribution, with maximum and minimum possible values specified. This code would enable repeating the model several hundred or thousands of times with relative ease, and potentially remove the error message, for future assessments.

The 95% confidence intervals for each parameter estimated were obtained using the mxCI command within openMx. This command allows confidence intervals of a specified width to be estimated for any of the freely estimated parameters. These are determined through an iterative -2 log likelihood process that occurs after the parameter estimates have been achieved; each parameter value is increased until a -2 log likelihood is reached that is 95% greater than the starting value to estimate the upper limit, and the parameter value is decreased until a 95% larger -2 log likelihood is reached (35).

3.0 RESULTS

3.1 DESCRIPTION OF THE DATA

The observed relative frequencies of each larval stage are shown in Table 5 for raters one through seven.

Table 5. Observed relative frequencies

	Rater 1 (AAF)	Rater 2 (ALF)	Rater 3 (AGS)	Rater 4 (DAH)	Rater 5 (HNW)	Rater 6 (SAK)	Rater 7 (UMA)
L1	0.2	0.27	0.28	0.36	0.3	0.28	0.25
L2	0.2	0.13	0.18	0.17	0.13	0.15	0.18
dauer	0.2	0.23	0.15	0.1	0.15	0.18	0.15
L3	0.1	0.1	0.12	0.08	0.15	0.2	0.13
L4	0.3	0.27	0.27	0.29	0.27	0.18	0.28

For the L1 stage, the highest and lowest frequencies were 0.36 and 0.2, for L2 these were 0.13 and 0.2, 0.10 and 0.23 for dauer, 0.08 and 0.2 for L3, and 0.18 and 0.3 for L4. Hence, the largest discrepancy was in the L1 stage, between raters AAF and DAH. Figure 7 shows histograms for each rater, based on the numbers tabulated in Table 5. These graphs demonstrate that each rater is unique in how they judge this population for stage of development. Additionally, it appears that raters AGS, DAH, UMA, and HNW tend to place a larger fraction of worms in the extreme larval stages (L1 or L4)

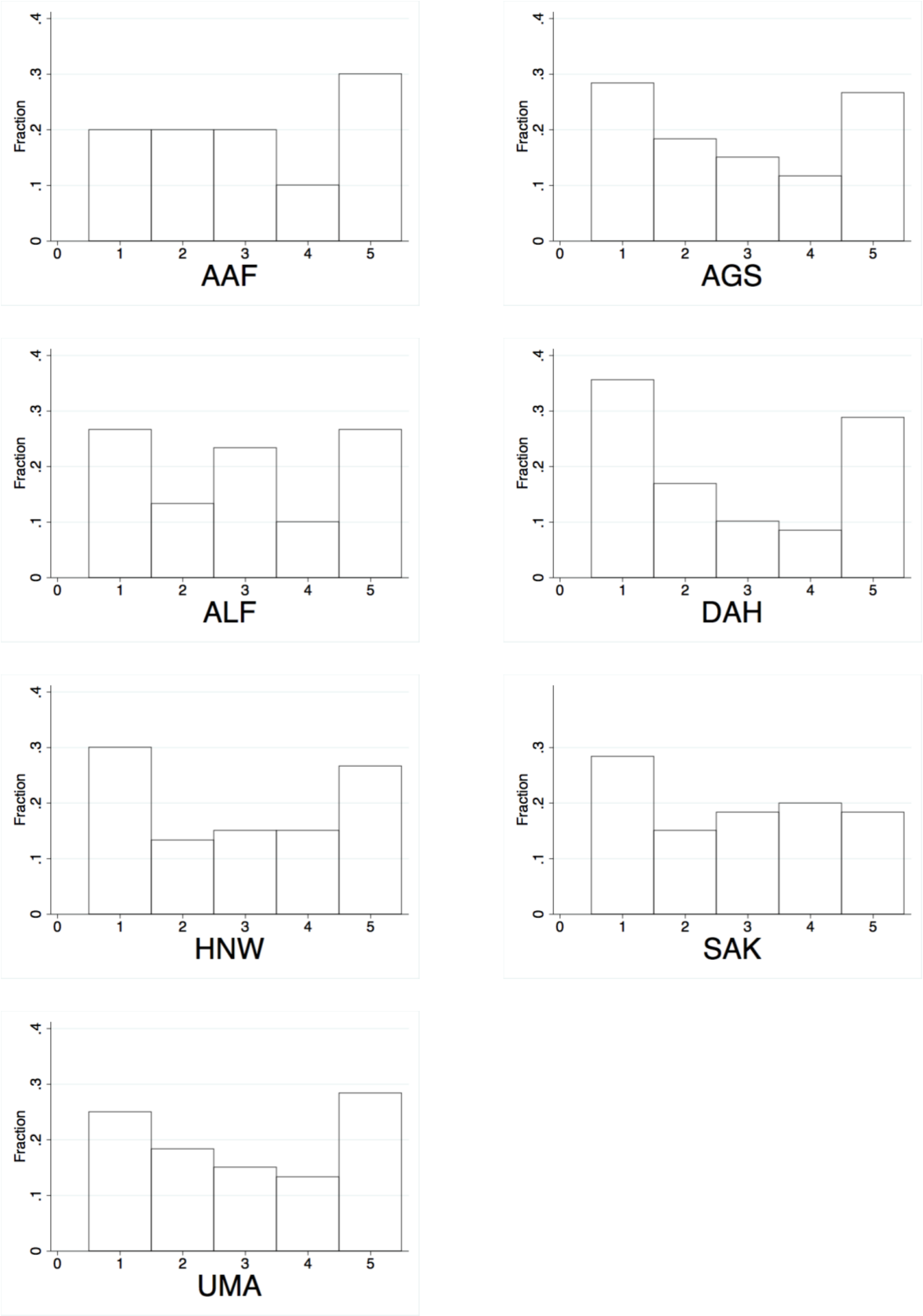

Figure 7. Histograms for each rater

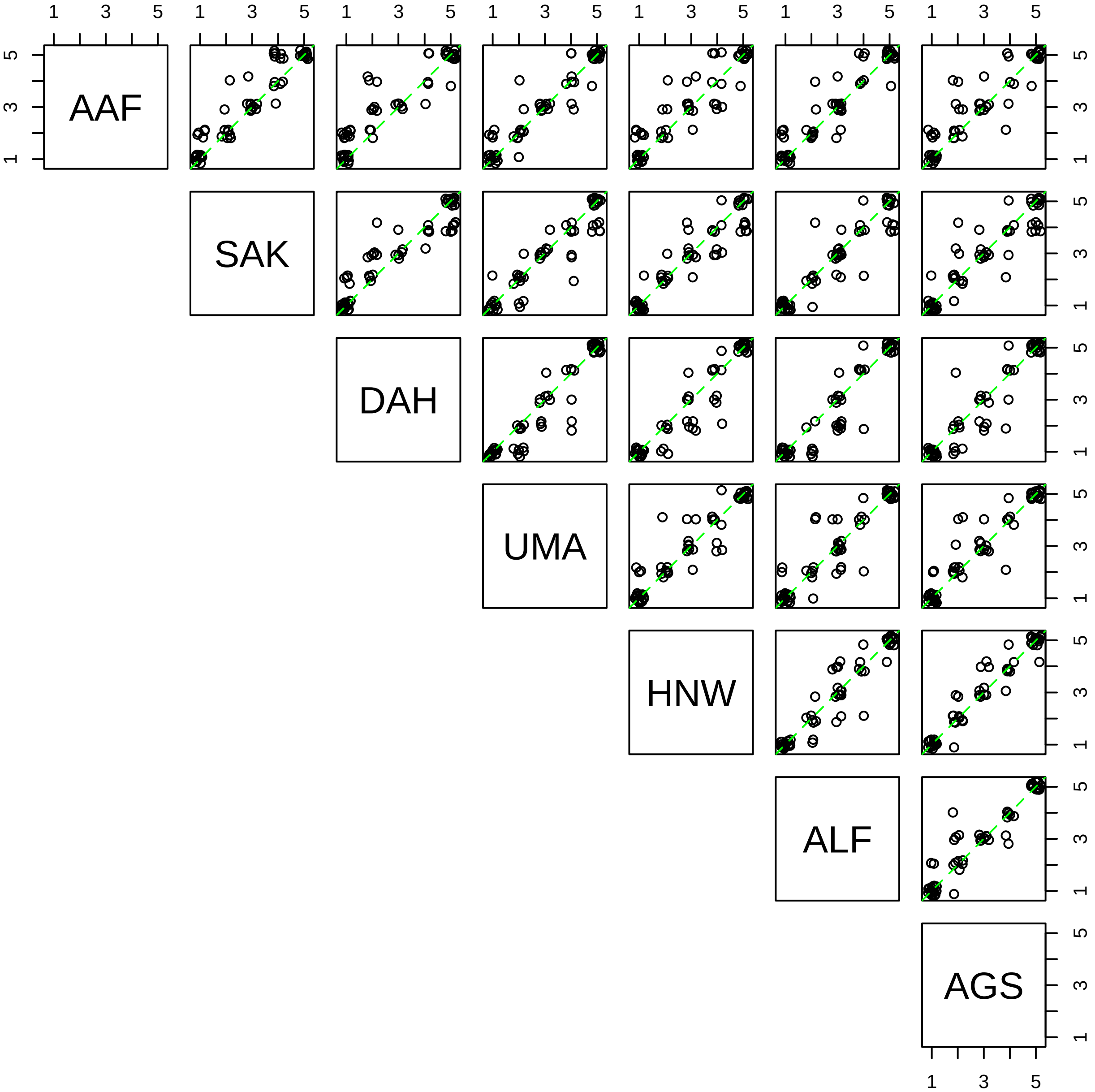

Figure 8. Pairwise error plots for comparison between raters

Figure 8 shows plots comparing the results between each rater. These pairwise error plots represent the results of one rater on the x-axis vs. another rater on the y-axis. In order to allow for visualization of overlapping points, a jitter function was used to add a small amount of noise to the data. Any points falling outside of the diagonal correspond to discrepancies between the two raters. For all of the plots, it appears that most of the point density is along the diagonal, indicating a considerable degree of agreement overall. The points tend to agree the most for the extreme categories (L1 and L4), and most of the discrepancies appear in the middle categories (L2, L3, and dauer). These plots also demonstrate that for each pair of raters, there is variability over whether the disagreements fall above or below the diagonal. For the pairs of raters with most of the discordant points lying in one direction of the diagonal, this is suggestive of a bias, or directionality of one rater compared to the other. For instance, in the plot of rater HNW vs. ALF, the points appear to fall in a relatively uniform distribution above or below the diagonal line, while in the plot of raters UMA vs. AGS, most of the discordant points lie above the diagonal line, suggesting a tendency of rater UMA to rank the subjects in higher categories compared to rater AGS.

3.2 ASSESSMENT OF INTERRATER RELIABILITY WITH THE KAPPA STATISTIC

As a preliminary assessment of interrater reliability between pairs of raters, the observed agreements are shown in Table 6. These values ranged from 66.1% to 86.67% concordant ratings. These pairwise ratings ranged from 0.5641 to 0.8295, corresponding to raters AAF and

DAH having the least agreement, and raters UMA and AGS having the highest agreement. Most of these values have a "substantial" strength of agreement (kappa between 0.61 and 0.8), with one pair having a "moderate" (0.41 – 0.6) strength kappa, and three pairs with "almost perfect" (0.81 – 1) (3, 4, 5)

Table 6. Pairwise observed and expected agreement and kappa values, equal ratings.

	Observed Agreement	Expected Agreement	Kappa	Strength of agreement (Landis)
AAF vs. SAK	73.33%	19.83%	0.6674	substantial
AAF vs. DAH	66.10%	22.23%	0.5641	moderate
AAF vs. UMA	80.00%	21.50%	0.7452	substantial
AAF vs. HNW	66.67%	21.17%	0.5772	moderate
AAF vs. ALF	78.33%	21.67%	0.7234	substantial
AAF vs. AGS	71.67%	21.50%	0.6391	substantial
SAK vs. DAH	69.49%	21.55%	0.6111	substantial
SAK vs. UMA	75.00%	20.44%	0.6858	substantial
SAK vs. HNW	73.33%	21.14%	0.6619	substantial
SAK vs. ALF	78.33%	20.72%	0.7267	substantial
SAK vs. AGS	75.00%	20.75%	0.6845	substantial
DAH vs. UMA	77.97%	23.07%	0.7136	substantial
DAH vs. HNW	76.27%	23.53%	0.6897	substantial
DAH vs. ALF	74.58%	22.75%	0.6709	substantial
DAH vs. AGS	79.66%	23.50%	0.7341	substantial
UMA vs. HNW	81.67%	21.75%	0.7657	substantial
UMA vs. ALF	80.00%	21.50%	0.7452	substantial
UMA vs. AGS	86.67%	21.81%	0.8295	almost perfect
HNW vs. ALF	80.00%	21.89%	0.7440	substantial
HNW vs. AGS	85.00%	22.06%	0.8076	almost perfect
ALF vs. AGS	85.00%	21.78%	0.8082	almost perfect

To assess the kappa values across all seven raters, kappa values were calculated for each rating category, and an overall kappa values was calculated; these results are displayed in Table 7. There is the least amount of agreement across raters for the L3 and L2 categories (with "fair" and "moderate" strength of agreement, respectively); the dauer stage had a "moderate" strength of agreement, and the L1 and L4 stages had "almost perfect" agreement. These ratings agree with a visual assessment of the pairwise error plots, which show a common pattern between raters, of many points falling off the diagonal for the middle categories, and few doing so in the extreme categories. The combined kappa value is considered "substantial" agreement.

Table 7. Multiple rater kappa

Outcome	Kappa
L1	0.8450
L2	0.5102
dauer	0.6745
L3	0.4460
L4	0.8610
combined	0.7034

3.3 ASSESSMENT OF INTERRATER RELIABILITY WITH LATENT VARIABLE MODELING

Figure 9 shows the one factor ordinal model that was fit to this data. The latent variables are outlined with circles, with variable S representing the latent common factor, which is the continuous distribution of the population stage of development, and variables Y_1, through Y_7 representing the underlying continuous distribution along which each rater assesses the stage of development. In this model, variation in the common factor is independently explained by each of the rater's underlying distribtion (as indicated by the single-headed arrows pointing to each of the Y variables), and by residual error (as indicated by the double headed arrow). The manifest varaibles are indicated by boxes are P_1 through P_7. These are each related to a continuous latent

Y variable by the probit function, which estimates threshold parameters along a standard normal distiribution. This relationship is indicated by the squiggly arrows connecting the manifest variables to the latents. The parameters in this model are the b_1 through b_7, which are the factor loadings that relate each rater to the common factor, as well as individual category threshold values for each rater (which are equivalent to the number of categories minus one), and residual error terms. Hence, there are seven factor loadings, 4 X 7 = 28 thresholds, and eight residual errors, for a total of 43 parameters. To make this model identified, it is necessary to fix certain parameters and variables. The distribution for the common factor is constrained to the standard normal distribution, with a variance of 1.0, and the distributions for each rater's latent variable are also set to the standard normal. Because of the constraints of the model, the factor loadings have been standardized, and their estimated values fall between -1 and 1. The threshold cut-off parameters for each rater are z-score values that lie along a standard normal distribution. Figure 10 depicts the threshold values along a standard normal distribution. The area under the curve between two threshold values represents the expected proportion of the particular stage of development a given rater will assign. With these constriants, the model has 35 freely estimated parameters.

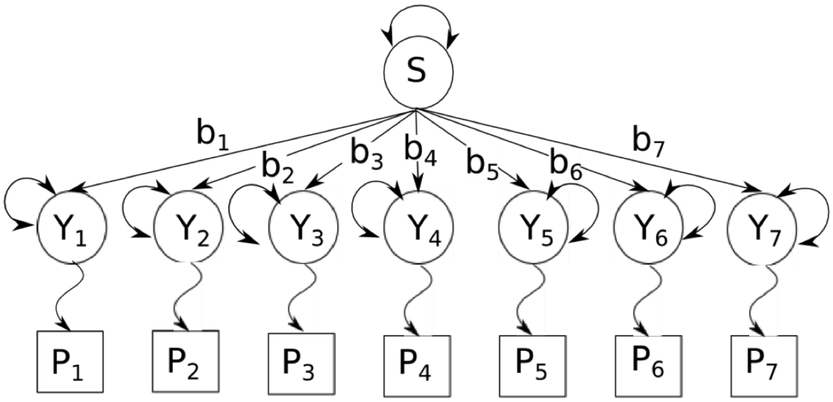

Figure 9. Path diagram for the one-factor ordinal model with 7 raters.

The full information maximum likelihood estimates for the factor loadings are shown in Table 8. These values are all very high, ranging from 0.982 to 0.999, meaning that all of the seven raters showed minimal error in predicting the common factor, given that a factor loading value of 1 would indicate no error, or perfect correlation to the common factor for a particular rater. Rater AAF had the lowest value of 0.982, while rater HNW had the highest value at 0.999, however, the factor loading values for all raters are all very close to one, and all have fairly narrow widths (the lowest lower bound was from rater AAF, at 0.971, which is still a high correlation). There is little indication from any of these values that the raters have any considerable lack of overall correlation with the common factor. To explain the apparent variation present in the data, the threshold values are more informative.

Table 8. Factor loading estimates for one factor ordinal model, with seven raters

Rater	parameter	estimate	95% C.I.	
			lower	upper
1 (AAF)	b_1	0.982	0.971	0.989
2 (ALF)	b_2	0.989	0.986	0.992
3 (AGS)	b_3	0.995	0.99	0.996
4 (DAH)	b_4	0.997	0.994	NA*
5 (HNW)	b_5	0.999	0.999	0.999
6 (SAK)	b_6	0.988	0.979	0.99
7 (UMA)	b_7	0.991	NA	0.993

*NA indicates instances where an estimate was not reached for the interval value

Table 9 displays the threshold values for all seven raters, where each threshold value represents a z-score along the standard normal curve. Because there are five possible categories for the rater to choose from (L1, L2, L3, dauer, or L4), the first threshold may be interpreted as the cuttoff value for the probability a rater will categorize as worm as an L1 (a z-score lower than the first threshold gets an L1 ranking, and higher means L2, L3, dauer or L4). While the z-score values themselves do not translate to concrete quantities for how a worm is rated, they are informative in comparing one rater to another and informing on whether there are specific categories over which raters tend to disagree the most.

Table 9. z-score threshold estimates for seven raters

Rater	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1 (AAF)	-0.897	-0.165	0.223	0.558
2 (ALF)	-0.658	-0.142	0.266	0.695
3 (AGS)	-0.605	0.005	0.208	0.718
4 (DAH)	-0.307	0.092	0.233	0.65
5 (HNW)	-0.538	-0.035	0.137	0.753
6 (SAK)	-0.594	-0.083	0.247	0.928
7 (UMA)	-0.725	-0.062	0.162	0.629

Table 10 has the expected frequencies for each stage of development and for each rater. These estimates are obtained by calculating the area under the curve of a standard normal distribution between each z-score. The expected frequency for the L1 stage was obtained by calculating the area under the curve from minus infinity to threshold 1, L2 was the area between threshold 1 and 2, dauer was between threshold 2 and 3, L3 was between thresholds 3 and 4, and L4 was the area from threshold 4 to infinity.

Table 10. Expected frequencies of larval stages for each rater

Rater	L1	L2	dauer	L3	L4
1 (AAF)	0.185	0.249	0.154	0.123	0.288
2 (ALF)	0.255	0.189	0.161	0.152	0.244
3 (AGS)	0.273	0.229	0.08	0.181	0.236
4 (DAH)	0.379	0.158	0.055	0.15	0.258
5 (HNW)	0.295	0.191	0.068	0.22	0.226
6 (SAK)	0.276	0.191	0.131	0.226	0.177
7 (UMA)	0.234	0.241	0.089	0.171	0.265
Average	0.27	0.21	0.11	0.17	0.24

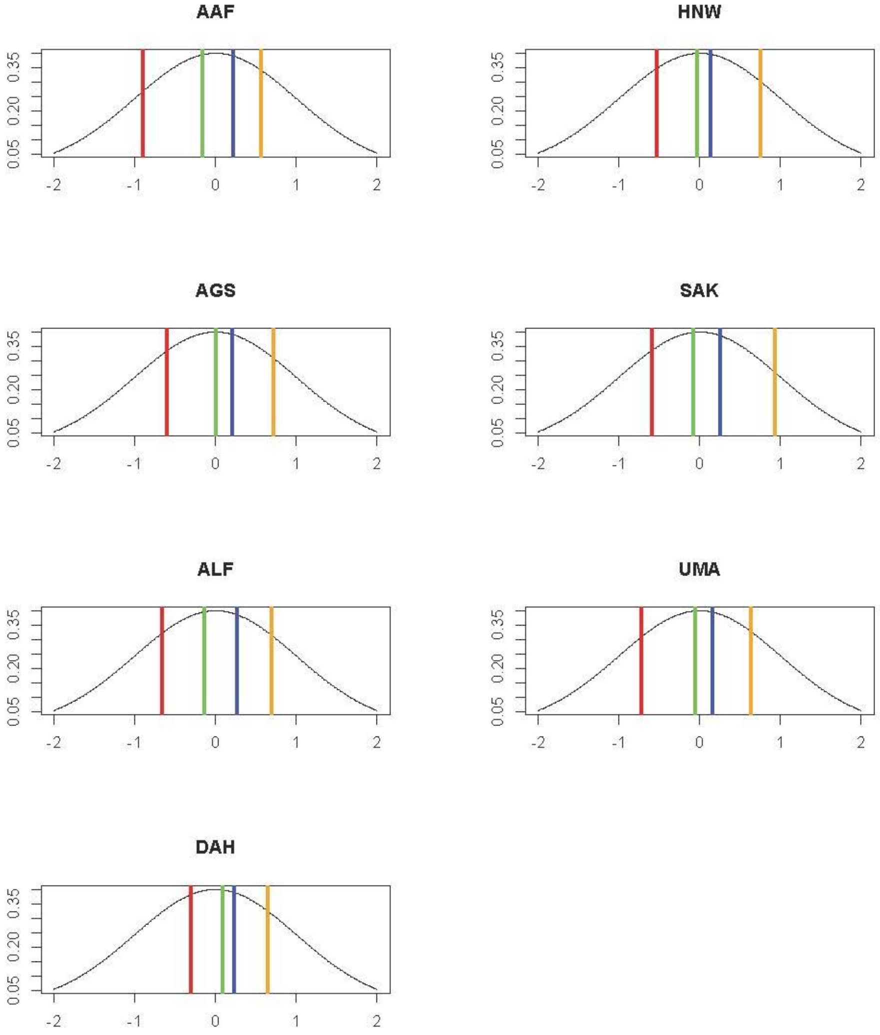

Figure 10. Estimated z-score threshold values for each rater red=threshold 1, green=threshold 2, blue=threshold 3, and orange=threshold4

Figure 10 displays the relative locations of each rater's threshold values along the x-axis of a standard normal distribution, to recapitulate the information contained in Tables 9 and 10. Rater AAF seems to have a considerably lower value for threshold 1 compared to the others, while raters HNW and DAH seem to have much smaller areas between thresholds 2 and 3, which corresponds to the frequency of dauer.

To give a visual assessment of model fit, Figure 10 displays plots of the expected probabilities of each stage of development for each rater, shown next to plots for the observed frequencies for these stages. Comparing these plots by eye, the general trend for each observed plot is present in the predicted plot; specifically raters ALF, AGS, DAH, and HNW have higher observed frequencies in the L1 and L4 stages relative to the L2, dauer and L3 stages, and similarly, the expected frequency graphs show this pattern. Rater SAK has a high observed proportion in the L1 stage relative to the other stages, while the predicted graph shows a similar pattern, and raters AAF and UMA both have higher relative proportions in the L4 stage, compared to the other stages in both the observed and the expected graphs. The graphs display slight discrepancies between the predicted and observed frequencies for dauer; these appear to be lower for all the raters' predicted frequencies compared to the observed. These graphs do not reveal any glaring difference in trend between the observed and expected, and are indicative of a fairly good model fit. Table 11 recapitulates the information in Figure 11, but provides numerical values to compare observed to expected relative frequencies side by side.

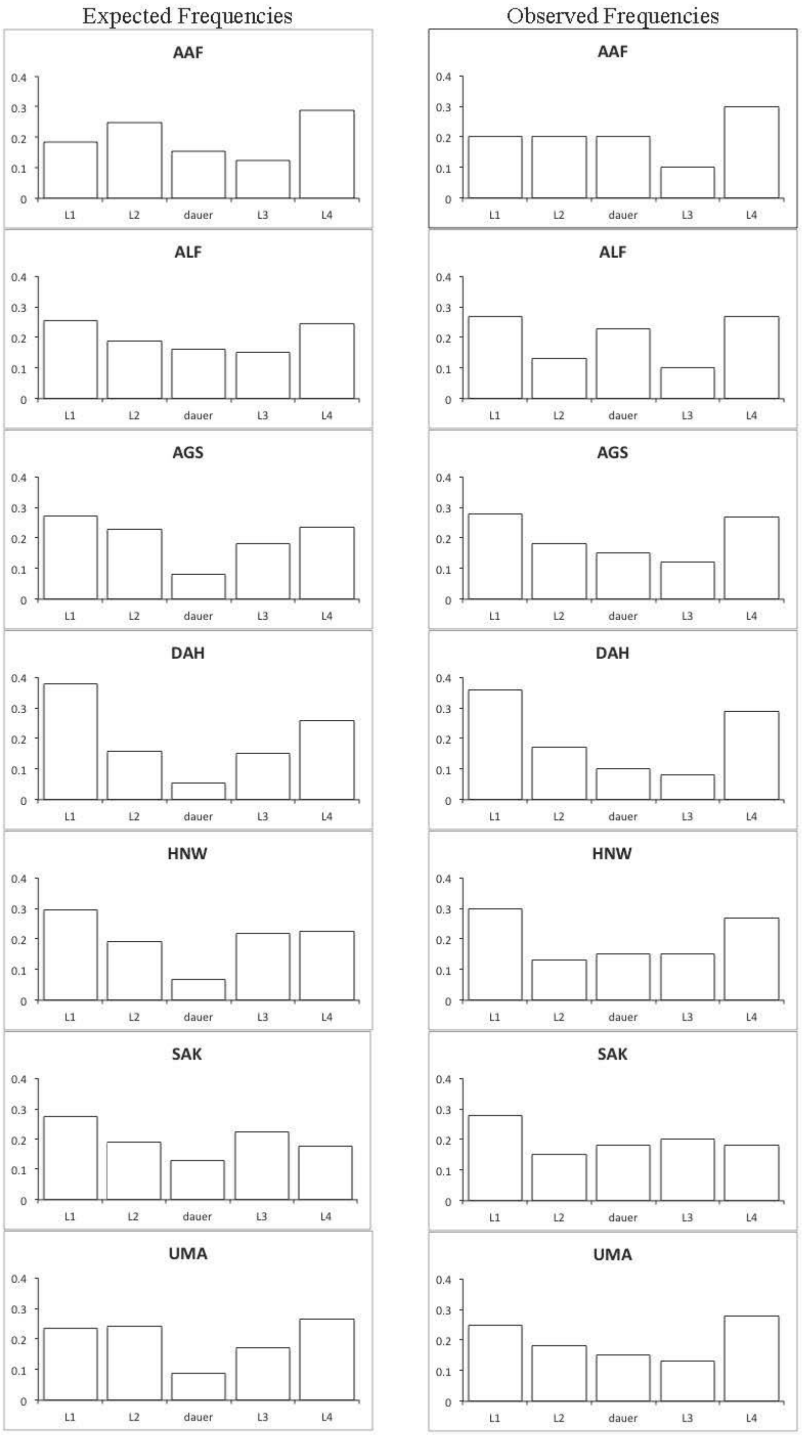

Figure 11. Plots showing predicted and observed developmental stage frequencies for each rater.

Table 11. Observed and expected frequencies of each stage

Rater	L1	L2	dauer	L3	L4
1 (AAF) exp	0.185	0.249	0.154	0.123	0.288
1 (AAF) obs	0.2	0.2	0.2	0.1	0.3
2 (ALF) exp	0.255	0.189	0.161	0.152	0.244
2 (ALF) obs	0.27	0.13	0.23	0.1	0.27
3 (AGS) exp	0.273	0.229	0.08	0.181	0.236
3 (AGS) obs	0.28	0.18	0.15	0.12	0.27
4 (DAH) exp	0.379	0.158	0.055	0.15	0.258
4 (DAH) obs	0.36	0.17	0.1	0.08	0.29
5 (HNW) exp	0.295	0.191	0.068	0.22	0.226
5 (HNW) obs	0.3	0.13	0.15	0.15	0.27
6 (SAK) exp	0.276	0.191	0.131	0.226	0.177
6 (SAK) obs	0.28	0.15	0.18	0.2	0.18
7 (UMA) exp	0.234	0.241	0.089	0.171	0.265
7 (UMA) obs	0.25	0.18	0.25	0.13	0.28

3.3.1 Hypothesis Testing

In order to formally test the null hypothesis that particular category thresholds are the same as others, Table 12 displays each parameter, with its 95% C.I.

Table 12. Threshold point and interval estimates

	Rater	lower bound	estimate	upper bound
Threshold 1	1 (AAF)	-1.264	-0.897	NA*
	2 (ALF)	-0.772	-0.658	-0.498
	3 (AGS)	-0.698	-0.605	-0.534
	4 (DAH)	-0.571	-0.307	-0.235
	5 (HNW)	-0.601	-0.538	-0.444
	6 (SAK)	-0.909	-0.594	-0.315
	7 (UMA)	-0.944	-0.725	-0.457
Threshold 2	1 (AAF)	-0.172	-0.165	0.067
	2 (ALF)	-0.175	-0.142	0.019
	3 (AGS)	-0.146	0.005	0.255
	4 (DAH)	-0.055	0.092	0.254
	5 (HNW)	-0.098	-0.035	0.154
	6 (SAK)	-0.154	-0.083	0.123
	7 (UMA)	NA*	-0.062	0.189
Threshold 3	1 (AAF)	0.077	0.223	0.401
	2 (ALF)	0.207	0.266	0.458
	3 (AGS)	0.185	0.192	0.192
	4 (DAH)	0.165	0.233	0.287
	5 (HNW)	0.096	0.137	0.238
	6 (SAK)	0.151	0.247	0.298
	7 (UMA)	0.086	0.162	0.234
Threshold 4	1 (AAF)	0.421	0.558	0.612
	2 (ALF)	0.545	0.695	0.799
	3 (AGS)	0.521	0.702	0.747
	4 (DAH)	0.526	0.65	0.743
	5 (HNW)	0.557	0.753	0.87
	6 (SAK)	NA*	0.928	1.151
	7 (UMA)	0.543	0.629	0.692

* NA indicates instances where an estimate was not reached for the interval value

Figure 12 displays the graphs with the relative 95% confidence interval widths, which allows for an easy visual determination of any overlap present. Separate plots were made for each threshold, and the x-axis designates the raters one through seven.

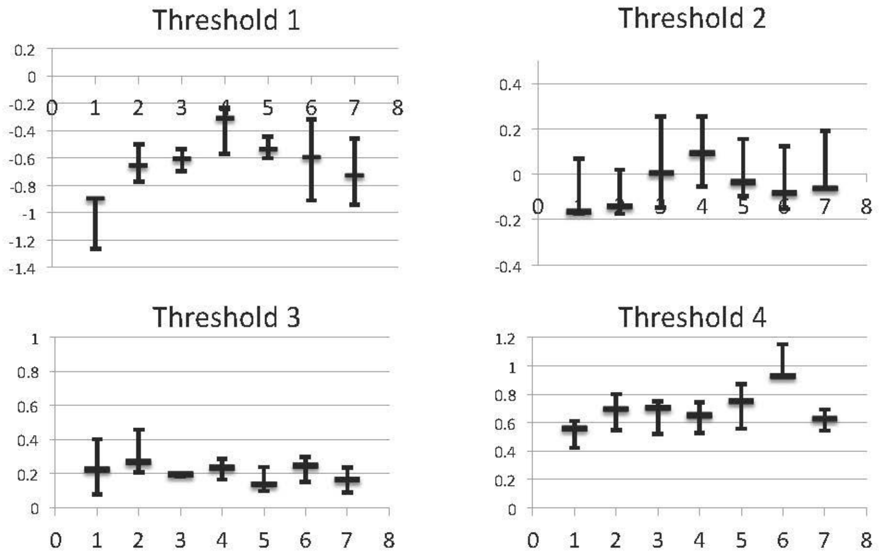

Figure 12. Plots showing the relative confidence interval widths for each rater's threshold values. (1=AAF, 2=ALF, 3=AGS, 4=DAH, 5=HNW, 6=SAK, 7=UMA)

To test whether any one of the raters has significantly different threshold values from rater ALF, Table 13 displays whether or not there is overlap in the confidence intervals.

Table 13. 95% Confidence interval overlap between ALF vs. other raters

Rater	Threshold 1	Threshold 2	Threshold 3	Threshold 4
1 (AAF)	no	yes	yes	yes
3 (AGS)	yes	yes	no	yes
4 (DAH)	yes	yes	yes	yes
5 (HNW)	yes	yes	yes	yes
6 (SAK)	yes	yes	yes	no
7(UMA)	yes	yes	yes	yes

Almost all of the confidence intervals for thresholds 1 through 4 show some degree of overlap, meaning that this model doesn't show evidence for a lack of agreement between the thresholds for ALF compared to those of the other raters. Rater AAF does show disagreement with rater ALF for the first threshold; the point estimate for AAF's threshold is -0.897, while that of ALF is -0.658. Rater AAF's threshold 1 estimate is considerably lower than the other rater's threshold 1 estimates as well (point estimates of -0.605, -0.307, -0.538, -0.594, -0.725 for raters AGS, DAH, HNW, SAK, and UMA, respectively), and rater AAF's interval estimate did not overlap with the interval estimates of raters AGS, DAH, and HNW.

Table 14. Larval stage frequency estimates and 95% Confidence interval for each rater.

	Rater	lower bound	estimate	upper bound
L1	1 (AAF)	0.103	0.185	NA
	2 (ALF)	0.220	0.273	0.309
	3 (AGS)	0.243	0.273	0.297
	4 (DAH)	0.284	0.379	0.407
	5 (HNW)	0.274	0.295	0.329
	6 (SAK)	0.182	0.276	0.376
	7 (UMA)	0.173	0.234	0.324
L2	1 (AAF)	NA*	0.250	0.424
	2 (ALF)	0.121	0.171	0.288
	3 (AGS)	0.145	0.229	0.358
	4 (DAH)	0.071	0.157	0.316
	5 (HNW)	0.132	0.191	0.287
	6 (SAK)	0.062	0.191	0.367
	7 (UMA)	NA*	0.241	0.402
dauer	1 (AAF)	0.004	0.154	0.224
	2 (ALF)	0.074	0.161	0.246
	3 (AGS)	0.000	0.074	0.134
	4 (DAH)	0.000	0.055	0.135
	5 (HNW)	0.000	0.068	0.133
	6 (SAK)	0.011	0.131	0.178
	7 (UMA)	-0.041	0.089	NA*
L3	1 (AAF)	0.007	0.123	0.199
	2 (ALF)	0.031	0.152	1.370
	3 (AGS)	0.123	0.183	0.199
	4 (DAH)	0.088	0.150	0.206
	5 (HNW)	0.117	0.220	0.270
	6 (SAK)	NA*	0.226	0.315
	7 (UMA)	0.114	0.171	0.221
L4	1 (AAF)	0.270	0.288	0.337
	2 (ALF)	0.212	0.244	0.293
	3 (AGS)	0.228	0.241	0.301
	4 (DAH)	0.229	0.258	0.299
	5 (HNW)	0.192	0.226	0.289
	6 (SAK)	0.125	0.177	NA*
	7 (UMA)	0.244	0.265	0.294

* NA indicates instances where an estimate was not reached for the interval value

Table 14 shows the point estimates and 95% confidence interval estimates converted from the threshold estimates, to give the areas under the standard normal distribution between two thresholds. The lower and upper bounds represent the largest and smallest possible areas between thresholds that would be given by the lower and upper confidence limits for the threshold estimates.

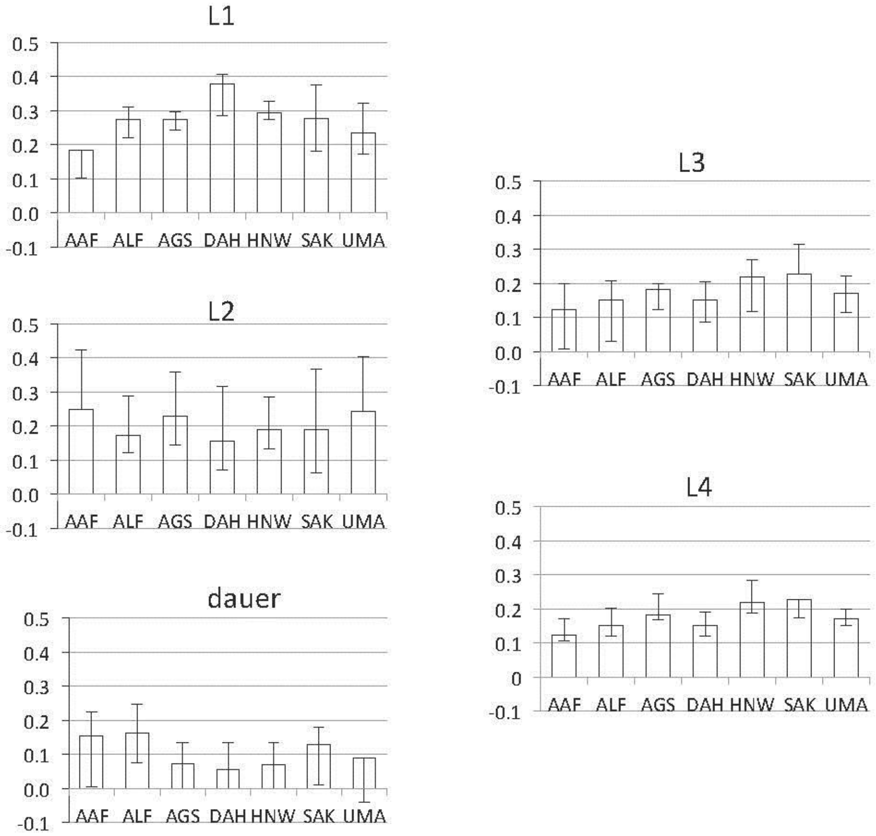

Figure 13. Estimated larval frequency and 95% confidence interval for each rater.

Figure 13 displays the information from Table14 in a graphical form. Each graph represents a different stage of development, with the y-axis as the frequency, and a bar for each rater. The L2 and dauer stages appear to have the widest confidence intervals, while the L2 stage as relatively narrow confidence intervals, but shows differences in point estimates from rater to rater. The L4 stage appears to have the least amount of variability, based on the narrow confidence intervals, and the similarity of the point estimates, and the L3 stage has fairly close point estimates, but wider confidence intervals.

4.0 DISCUSSION

The aims of this project were to design and carry out an experiment for measuring interrater reliability among *C. elegans* researchers, to evaluate these data using both the kappa statistic and a one-factor ordinal latent variable model, and to draw a conclusion about which statistical method is the most appropriate and useful for this dataset. The major findings were that 1) it is possible to quickly obtain data from multiple raters assessing the same population of worms for five stages of development, 2) that visual assessment of pairwise error plots reveals some amount of error and bias among the seven raters, especially for the L2, dauer and L3 stages, 3) that the overall kappa value is fairly high, but that kappa showed worse agreement for the L2 and L3 stages compared to the other stages, and 4) that in the one factor model, all raters have very low random error, but they vary in their threshold estimates.

The goal of the experimental design was to minimize any sources of variability in the data that were not due to rater error. To ensure that each rater was making an assessment under the same visual conditions, a video recording of a magnified population of worms was used instead of having the raters look at live worms under a microscope. This enabled each worm to be given a unique identifier to eliminate the possibility that raters were assessing a different set of worms, or that the same worm was rated twice. One criticism of this design is that by using a pre-recorded image, this is taking away certain qualities of a visual assessment a rater may use when looking under a microscope, such as adjusting the zoom and focus of the lens; these adjustments might be essential in a rater's ability to make a judgment, and this design does not evaluate this aspect of how a rater reaches this judgment. However, to gain consistency in the population being assessed, and for the feasibility of gathering data from multiple raters, it was assumed that having each rater view a movie was similar to what a researcher would do to make an assessment in the field.

To visualize disagreements in this dataset pairwise error plots were generated (Figure 8) in which each rater's results were plotted against those of the other raters. The points that lie along the diagonal line for each plot represent agreement between raters, while those that fall off the diagonal are discordant ratings. These graphs reveal that overall most of the points are concordant. They also provide a means of visualizing the degree of discordance between points, and whether there are biases or patterns in the disagreements. In all of the graphs, most of the discordant points fall closer to the center of the graph, meaning that most of the discrepancies arise from classifications in the L2, L3 and dauer stages. The pairwise error plots also exhibited certain amounts of bias between some of the raters, based on a majority of discordant points lying either above or below the diagonal line. By evaluating the error plots, it appears that rater DAH seems to usually give lower ratings compared to other raters. In comparing UMA vs. AGS, UMA seems to rank worms higher. Additionally, to evaluate rater ALF, who has the most experience in *C. elegans* research, there is not striking visual for a strong bias in most of the pairwise plots. There is an apparent slight bias in ALF vs. AGS and ALF vs. AAF for ALF to rank worms higher in both cases.

While the pairwise error plots provide visual information about the overall strength of agreement, the bias of one rater over another, and about any tendency for one or more categories to have higher disagreement, they do not provide a formal analysis or a numerical index for the agreement. The kappa results give numerical values for the degree of agreement between these raters. The overall kappa value was 0.70, which is considered substantial agreement, and which corroborates a visual assessment of the pairwise plots that most points fall along the diagonal line. In addition, the stage-wise multi-rater kappa values show a similar result as the pairwise error plots: the L1 and L4 stages had high kappa values of 0.85 and 0.86, respectively, while the L2, dauer and L3 stages had slightly lower kappas of 0.51, 0.67, and 0.45 for L2, dauer and L3, respectively, indicating less agreement for the L2, dauer, and L3 stages compared to the L1 and L4 stages.

For this type of data, the use of kappa as a measure of interrater reliability is not ideal. The kappa calculations used in this analysis were devised to contrast raters using a categorical scale, where the rating categories do not have a natural ordering(3). In this dataset, because of the ordering of the stages of development based on worm length and thickness (L1, L2, dauer, L3, and L4), if one rater classifies a worm as an L2 while another rater calls the same worm an L4, this is a more pronounced disagreement than L1 vs. L2. The kappa statistic assumes that all instances of discordance are the same in severity. A weighted kappa statistic that accounts for an ordering in the categories exists. However, the decision for what weight to give each category requires some guesswork. For this kappa analysis, the categories were all given the same weight. Therefore, the kappa results may be somewhat conservative, given that disagreements between adjacent categories have the same magnitude as those between non-adjacent categories and a majority of the disagreements within this set are between adjacent categories.

Additional problems arise from the kappa statistic, as were discussed in the introduction. These include the possibility of the kappa value giving an inaccurate representation of agreement due to the prevalence of the different larval stages in the population. This is unlikely to be a major issue for this dataset, given that there was not much sparseness for any of the larval stages in this populations (the lowest average observed frequency across seven raters was 13%, for the L3 stage). The other issue with kappa is that it fails to provide information about bias or directionality of the disagreements. While the kappa value indicates the strength of overall agreement, it doesn't provide an explanation of bias or directionality of disagreement; the kappa value contains both elements of rater bias and error without allowing for a separate consideration each. Evaluating the data in graphical representations, and by fitting a latent variable model allows for concerns over rater bias and the use of ordered categorical outcomes to be accounted for.

The model that was fit to this data had a total of 35 freely estimated parameters, which each had an estimated 95% confidence interval. As a result, there are a large number of formal comparisons between raters that may be done to draw conclusions about the interrater reliability in this dataset (including pairwise comparisons between raters, and comparisons of one rater versus the group of raters). The estimation of both factor loading parameters and threshold parameters provides two different types of error measurement and comparison between raters. The threshold values provide distinctions for how each rater categorizes the worms, and gives information about rater biases, or tendencies to rate worms in one category over the others. Factor loadings, on the other hand, give an indication of the amount of random error each rater has.

In this model, the factor loadings were very close to one, meaning that all of the raters had a high correlation to the group average, and that the amount of random error that was not accounted for in the threshold differences is very low. Furthermore, there were not any raters who stood out in terms of having a low factor loading compared to the others. Therefore, most of the variability present between the raters was accounted for by differences in the threshold values.

While a large number of comparisons are possible for the threshold estimates from this model, there were a few that were of more interest for the focus of this study. Specifically, 1) assessments were made to determine whether differences in threshold values explain the apparant biases that are observed in the error plots, 2) comparisons were made between each rater and rater ALF, and 3) comparisons were made to determine whether any rater has a different definition for the dauer stage. One aim of making these comparisons was to determine how the raters might adjust their rating strategy to improve their agreement with the other raters

Comparing the threshold values from rater to rater gives a determination of differences in category definitions between raters, as well as biases that are visually apparent in the pairwise error plots. If the estimated threshold values are considerably different between raters, this would indicate a tendency of one rater to classify worms in a higher category than the other. Since rater DAH appears to have a bias towards lower ratings, one would expect the threshold values to be higher than the other raters. Indeed, DAH has the highest values for thresholds 1 and 2 (-0.307, for threshold 1, compared to less than -0.53 for the other raters, and 0.092 for threshold 2, compared to 0.005 or less for the other raters). DAH has a higher expected frequency in the L1 stage compared to the others (0.379 vs. 0.295 or less), and a roughly comparable L2 expected frequency, which further indicates that DAH has incorrectly classified some of the higher staged worms as L1s. Based on these results, to improve the reliability among these raters, rater DAH ought to re-evaluate how to distinguish the younger staged worms from the older ones, and be informed of his tendency to rank worms lower. Based on threshold values, rater SAK has a much higher threshold 4 value, compared to the other raters (the 95% confidence intervals do not overlap with those of the other raters), and rater AAF has a considerably lower threshold 1 value (the 95% C.I.'s fail to overlap with all other raters, with the exception of SAK). This information would suggest that rater SAK tends to miss worms that have reached the L4 stage, and that rater AAF tends to rank L1 worms into higher categories.

Another formal assessment one can make from the threshold estimates is to decide whether any particular rater is showing a lack of agreement on the dauer stage. The correct identification of dauer larvae is particularly important in *C. elegans* research because many studies investigate the inappropriate entry into dauer. Any differences between raters for threshold 2 or threshold 3 would indicate some amount of disagreement in dauer classification. For this dataset, the point estimate values for threshold two range from -0.165 to 0.092, and none of the individual raters have a lack of overlap in confidence intervals compared to the other raters. Hence this model doesn't provide evidence for any one of the raters having a difference in distinguishing the dauer stage from the L1 or L2 stages. For threshold 3, raters ALF and AGS have a lack of interval overlap, as was discussed previously, while none of the other raters have a lack of overlap. For both thresholds 2 and 3, there is no evidence of any of the individual raters having an interval estimate that fails to overlap with all of the other raters. Hence this model indicates that these raters have a good amount of agreement on threshold designations for the dauer stage.

While there are some statistically significant discrepancies in threshold values between some raters, there were not any specific raters who stood out as being grossly different from the others by having consistently higher or lower category definitions, or by having a considerably lower factor loading value. Since rater ALF had the most experience and expertise in the worm field, a reasonable set of null hypotheses to test are that rater ALF's thresholds are the same as each of the other raters'. A comparison of each rater to ALF indicated that the only raters with significant differences in threshold definitions were rater AAF, who had a much lower value threshold 1, and rater SAK for threshold 4.

While this model does provide information for how to minutely adjust how specific raters categorize worms, it does not give an indication of any of the raters classifying the worms in a considerably different way. Thus, the one factor ordinal model reveals minor areas of discrepancies between raters, while also conveying a general overall agreement between raters.

One question that arises from the results of these two analyses is to what degree do the two methods for interrater reliability assessment agree with each other. To begin with, it is not feasible to get a numerical value from the modeling parameter estimates that would be directly comparable with the kappa values. However, one can make a general overall conclusion based on the similarity of the factor loading values, and the threshold values; the fact that there is not a rater who stands out as having a dramatically different factor loading or threshold values means that the overall agreement is generally good. This conclusion certainly corroborates the overall multi-rater kappa value. Another interesting conclusion from the kappa value results for each stage of development was that the L1, dauer and L4 stages had high agreement, while the L2 and L3 stages did not. This conclusion fits with a visual assessment of the pairwise error plots. In order to make a similar assessment with the modeling results, the expected frequencies and 95% confidence intervals for each stage of development were assessed in Table 14 and Figure 12. Based on both the variability in point estimate and width of confidence intervals between raters in each graph, the L2 and dauer stages appear to have the most amount of disagreement, and the L4 stage has the best agreement between raters, with similar point estimates and narrow confidence intervals. While these graphs do not appear to clearly match the kappa statistics for these stages of development, the graphs contain information about the variability of the data, and the individual rater's predicted frequencies; the kappa value combines this information into one index.

There are obvious benefits to fitting a model to the data, however, this approach is much more computationally demanding than calculating the kappa statistic or doing simpler assessments like graphing the data and comparing proportions of agreement. Fitting this model becomes especially taxing as more parameters are involved. In this project, for instance, after over 100 attempts of altering the starting values, the software still provided a warning message with the parameter estimates, explaining that the lowest -2 log likelihood may have not been reached. After running the model several times, and achieving parameter estimates that were close in magnitude, the model with the lowest -2LL was selected for the final write-up. The automated code shown in Appendix D does minimize some of the labor involved in running the code multiple times, however, for this dataset achieving the best parameter estimates is lengthy process. To begin an attempt to achieve better estimates for this model for a future publication, the model was run 1000 times, to get a slightly low -2LL than shown in this report, with very similar parameter estimates. This indicates that the results reported in this draft are still not quite the best, but that they are very likely close enough parameter estimates to draw the same conclusions. Indeed, better estimates could be achieved after running the model more than 1000.

Despite the warning messages given from the modeling process, this does not diminish the reliability of the estimates. Running a model does, however take a lot more time and effort in comparison to calculating kappa statistics. A modeling approach is certainly worthwhile, if a detailed and informative assessment interrater reliability is desired; especially when one wishes be informed on how specific ways to improve the reliability of the raters. However, as preliminary easy assessments, the kappa and plotting the data may be sufficient.

5.0 CONCLUSION

An overall assessment of this data, based on the descriptive plots, percent agreement tables, the kappa statistic values, and the factor loading parameter estimates in the latent variable model is that there is a considerable amount of agreement between the raters. The lowest observed pairwise proportion of agreement was 66%, and the overall combined kappa value was 0.70, which is considered substantial agreement. The pairwise kappa values ranged from 0.564 between raters AAF and DAH, which is considered moderate, to 0.8296 between HNW and AGS, which is considered "almost perfect". The kappa values provided a general index of the strength of agreement, however, they failed to measure specific aspects of agreement such as rater bias. In this dataset, since there was a fairly strong amount of concordance between raters, the kappa values were at acceptable levels. In order to gain a better understanding of the sources of rater disagreement, the threshold values in the one factor ordinal model were more useful. Raters AAF and SAK stood out the most for having remarkably different threshold values compared to the other raters (specifically, threshold 1 for AAF and 4 for SAK). However, none of the raters differed significantly for all of the categories relative to the other raters.

The kappa statistic gave a general evaluation of overall reliability, however it did not provide useful information that coul help with improving specific aspects of worm categorization for particular raters.

APPENDIX A

SCORE CARD

Name ______________________________ Date _______________ Movie Number _________

years in a C. elegans lab: <1 _____ 1-3______ 3-6_____ 6-9_____ 9<_______

Worm #	L1	L2	dauer	L3	L4
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
...					

Worm#	L1	L2	dauer	L3	L4
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					
50					
51					
52					
...					

APPENDIX B

OPEN MX CODE: ONE FACTOR MODEL

Annotated R code (Ordinal one common factor model, 7 raters, 5 categories)

(adapted from openMx documentation: http://openmx.psyc.virginia.edu/docs/OpenMx/latest/FactorAnalysisOrdinal_Matrix.html)

1) Read in the raw data (7 raters, 5 rating categories indicated with values 1, 2, 3, 4 and 5)

```
        data_raw<-read.csv("/Volumes/NO       NAME/thesis       biostats/Movie_4_5categories.csv",
sep=",",header=TRUE)
        #remove worm_no
        data_raters<-data_raw[,c("AAF", "SAK", "DAH", "UMA", "HNW", "ALF", "AGS")]
```

2) Convert the numerical values to ordered factors using the MxFactor command

```
        ordinalData <- mxFactor(as.data.frame(data_raters),levels=c(1:(nThresholds+1)))
```

3) Label the number of raters (nVariables), number of latent factors (nFactors), and number of category thresholds (number of categories minus one, nThresholds), and rater names (raterNames) for the model.

```
        nVariables<-7
        nFactors<-1
        nThresholds<-4
        raterNames <- c("AAF", "SAK", "DAH", "UMA", "HNW", "ALF", "AGS")
```

4) To specify the model, first state a name for the model. Then generate a matrix to estimate the factor loadings, which has a row for each rater, and a single column (for the one factor). The Free=TRUE statement means that all the values will be freely estimated, and the lbound and ubound commands set the upper and lower limits of the factor loadings, which are standardized as a result of the common factor variance being fixed to one. The starting values for this matrix are given by the `values=` statement, which is set arbitrarily. This starting value may be adjusted if the program fails to converge on a minimum -2 log likelihood.

```
oneFactorThresholdModel <- mxModel("oneFactorThresholdModel",
        mxMatrix(
            type="Full",
            nrow=nVariables,
            ncol=nFactors,
            free=TRUE,
            values=0.96,
            lbound=-.999,
            ubound=.999,
            name="facLoadings"
        ),
```

5) To constrain the variances of the observed variables to one, a matrix consisting of one column for each rater, with a value of one. This is subtracted from the factor loadings squared. The squared factor loadings are obtained by multiplying the facLoadings matrix by the transposed facLoadings matrix, and applying the diag2vec function to this result.

```
       mxMatrix(
            type="Unit",
            nrow=nVariables,
            ncol=1,
            name="vectorofOnes"
        ),
       mxAlgebra(
           expression=vectorofOnes - (diag2vec(facLoadings %*% t(facLoadings))) ,
           name="resVariances"
       ),
```

6) The expected covariances are obtained by adding each of the residual variances to the square of the factor loadings.

```
       mxAlgebra(
            expression=facLoadings %*% t(facLoadings) + vec2diag(resVariances),
            name="expCovariances"
        ),
```

7) For the ordinal model, it is assumed that there is a latent distribution for each rater, with threshold values along this distribution that designate different categories. In this model, the latent distribution for each rater is constrained to the standard normal. Hence, the means for each rater are constricted to zero, while the threshold values are freely estimated. In the following statement, a zero vector is created, with columns for each rater.

```
       mxMatrix(
            type="Zero",
            nrow=1,
            ncol=nVariables,
            name="expMeans"
        ),
```

8) To estimate the threshold values for each rater, first a matrix containing threshold deviations is created, with a row for each threshold, and a column for each rater. These values are all freely estimated. The lower bounds for each estimate are set as minus infinity for the first threshold, and then 0.01 for each subsequent threshold deviation. Again, the starting values in the matrix following the `values=` statement may be adjusted if the model fails to converge.

```
   mxMatrix(
            type="Full",
            nrow=nThresholds,
            ncol=nVariables,
            free=TRUE,
            values=c(-1, 0.5, 0.5, 0.5),
            lbound=rep( c(-Inf,rep(.01,(nThresholds-1))) , nVariables),
            dimnames=list(c(), raterNames),
            name="thresholdDeviations"
        ),
```

9) These threshold deviations are converted to the expected thresholds by multiplying the thresholdDeviations matrix by a matrix with a lower half of ones.

```
mxMatrix(
    type="Lower",
    nrow=nThresholds,
    ncol=nThresholds,
    free=FALSE,
    values=1,
    name="unitLower"
),
mxAlgebra(
    expression=unitLower %*% thresholdDeviations,
    name="expThresholds"
),
```

10) The standard deviations are obtained by taking the square root of the variances. Additionally, likelihood based confidence intervals may be estimated for the threshold deviations.

```
 mxAlgebra(sqrt(resVariances),"resSDs")
 mxCI(c(‘‘thresholdDeviations’’)),
 mxCI(c(‘‘facLoadings’’))
)
```

11) Finally, the model is fit, using the `mxRun` command, and the output and model summary are obtained with the `@output` and `summary` commands.

```
oneFactorFit <- mxRun(oneFactorThresholdModel, intervals=TRUE)
oneFactorFit@output
summary(oneFactorFit)
```

APPENDIX C

OPEN MX CODE: RUNNING THE MODEL MANY TIMES

1) After the data is read-in, the number of thresholds and raters is stated, and the data is converted to ordinal data, the model is built as before.

```
m1 <- mxModel("m1",
       mxMatrix(
    type="Full",
    nrow=nVariables,
    ncol=nFactors,
    free=TRUE,
    values=0.96,
    lbound=-.999,
    ubound=.999,
    name="facLoadings"
       ),
       mxMatrix(
    type="Unit",
    nrow=nVariables,
    ncol=1,
    name="vectorofOnes"
       ),
       mxAlgebra(
    expression=vectorofOnes - (diag2vec(facLoadings %*% t(facLoadings))) ,
    name="resVariances"
       ),
       mxAlgebra(
    expression=facLoadings %*% t(facLoadings) + vec2diag(resVariances),
    name="expCovariances"
       ),
       mxMatrix(
    type="Zero",
    nrow=1,
    ncol=nVariables,
    name="expMeans"
       ),
       mxMatrix(
    type="Full",
    nrow=nThresholds,
    ncol=nVariables,
    free=TRUE,
    values=c(-1, 0.5, 0.5, 0.5),
    lbound=rep( c(-Inf,rep(.01,(nThresholds-1))) , nVariables),
    dimnames=list(c(), raterNames),
    name="thresholdDeviations"
       ),
       mxMatrix(
    type="Lower",
    nrow=nThresholds,
    ncol=nThresholds,
    free=FALSE,
    values=1,
    name="unitLower"
```

```
        ),
        mxAlgebra(
    expression=unitLower %*% thresholdDeviations,
    name="expThresholds"
        ),
        mxData(
    observed=ordinalData,
    type='raw'
        ),
        mxFIMLObjective(
    covariance="expCovariances",
    means="expMeans",
    dimnames=raterNames,
    thresholds="expThresholds"
        ), mxAlgebra(sqrt(resVariances), "resSDs")
   )
```

2) The following statements are required for fitting the model, and running it multiple times, with different starting values

3) First, the number of trials is stated:

```
trials <- 1000
```

4) Then the starting values for the parameter estimates are set up. These are set up as random values along the uniform distribution for each parameter estimate. For the factor loadings, the starting values are given between 0.8 and 0.99. For the threshold deviations, the first estimate is between -3 and 0, and the next three deviations fall between 0.1 and 0.5.

```
parNames <- names(omxGetParameters(m1))

input <- matrix(NA,trials,length(parNames))
dimnames(input) <- list(1:trials,parNames)

output <- matrix(NA,trials,length(parNames))
dimnames(output) <- list(1:trials,parNames)

fit <- matrix(NA,trials,4)
dimnames(fit) <- list(c(1:trials),c("Minus2LL","Status","Iterations","time"))

# Factor loadings
input[,"m1.facLoadings[1,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[2,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[3,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[4,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[5,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[6,1]"] <- runif(trials,0.8,0.99)
input[,"m1.facLoadings[7,1]"] <- runif(trials,0.8,0.99)

# Thresholds for rater 1
input[,"m1.thresholdDeviations[1,1]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,1]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,1]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,1]"] <- runif(trials,0.1,0.5)

# Thresholds for rater 2
input[,"m1.thresholdDeviations[1,2]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,2]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,2]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,2]"] <- runif(trials,0.1,0.5)

# Thresholds for rater 3
input[,"m1.thresholdDeviations[1,3]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,3]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,3]"] <- runif(trials,0.1,0.5)
```

```
input[,"m1.thresholdDeviations[4,3]"] <- runif(trials,0.1,0.5)


# Thresholds for rater 4
input[,"m1.thresholdDeviations[1,4]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,4]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,4]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,4]"] <- runif(trials,0.1,0.5)

# Thresholds for rater 5
input[,"m1.thresholdDeviations[1,5]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,5]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,5]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,5]"] <- runif(trials,0.1,0.5)

# Thresholds for rater 6
input[,"m1.thresholdDeviations[1,6]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,6]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,6]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,6]"] <- runif(trials,0.1,0.5)

# Thresholds for rater 7
input[,"m1.thresholdDeviations[1,7]"] <- runif(trials,-3,0)
input[,"m1.thresholdDeviations[2,7]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[3,7]"] <- runif(trials,0.1,0.5)
input[,"m1.thresholdDeviations[4,7]"] <- runif(trials,0.1,0.5)
```

5) A loop is set up that will fit the model the number of times specified in the trials statement, and will give starting values for the parameter estimates according to the input statements above.

```
# Loop to fit models

for(i in 1:trials) {

  temp1 <- omxSetParameters(m1,
             labels=parNames,
      values=input[i,]
      )

  temp1@name <- paste("Starting Values Set",i)

  temp2 <- mxRun(temp1,unsafe=TRUE,suppressWarnings=TRUE)

  output[i,] <- omxGetParameters(temp2)

  fit[i,] <- c(
    temp2@output$Minus2LogLikelihood,
    temp2@output$status[[1]],
    temp2@output$iterations,
    temp2@output$wallTime
    )

  print(output[i,])
  print(fit[i,])
  print(head(table(round(fit[,1],3),fit[,2])))

}

save.image(file="/Users/AAF/Desktop/AAFresults2.RData")
```

APPENDIX D

MODEL OPTIMIZATION

	starting values			parameter estimates				
no.	factor loadings	threshold deviations	-2 Log Likelihood	factor loadings	threshold 1	threshold 2	threshold 3	threshold 4
1	0.95	-0.8, 0.5, 0.25, 0.55	525.845	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.995	AAF: -0.900 ALF: -0.650 AGS:-0.340 DAH: -0.780 HNW:-0.590 SAK:-0.720 UMA: -0.650	AAF: 0.77 ALF: 0.55 AGS: 0.42 DAH: 0.71 HNW: 0.54 SAK: 0.56 UMA:0.65	AAF: 0.40 ALF: 0.35 AGS: 0.14 DAH: 0.23 HNW:0.18 SAK: 0.45 UMA:0.21	AAF: 0.35 ALF: 0.70 AGS: 0.45 DAH: 0.49 HNW: 0.64 SAK: 0.41 UMA:0.54
2	0.9826788, 0.9877669, 0.9966496, 0.9912003, 0.999, 0.9893618, 0.9950743	-0.886823, 0.7247954, 0.3845787, 0.3255732, -0.5859195, 0.5046829, 0.3326593, 0.6584063, -0.3012942, 0.3948908, 0.1411279, 0.4025556, -0.7153064, 0.6561375, 0.2222375, 0.4554093, -0.5260471, 0.4934621, 0.1711622, 0.5976217, -0.6480305, 0.5088975, 0.4037059, 0.4175175, -0.5961011, 0.6028839, 0.2017089, 0.4960843	525.8586	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.890 ALF: -0.590 AGS:-0.310 DAH: -0.720 HNW: -0.530 SAK: -0.650 UMA: -0.600	AAF: 0.72 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.49 SAK: 0.51 UMA: 0.60	AAF: 0.40 ALF: 0.33 AGS: 0.14 DAH: 0.22 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.34 ALF: 0.68 AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
3	0.9819188, 0.9882382, 0.9966982, 0.9911327, 0.999, 0.9892152, 0.9948704	-0.8965172, 0.7695836, 0.3992047, 0.3532392, -0.6460816, 0.5472614, 0.3452207, 0.7006323, -0.3383669, 0.421621, 0.1402084, 0.4460048, -0.7832472, 0.7057979, 0.2312134, 0.4866034, -0.590711, 0.544264, 0.1768126, 0.6368714, -0.7226458, 0.5551665, 0.4538412, 0.4140419, -0.6547262, 0.6460972, 0.2081288, 0.5428632	525.8587	AAF: 0.980 ALF: 0.990 AGS: 0.9970 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.89 ALF: -0.59 AGS: -0.31 DAH: -0.72 HNW: -0.53 SAK: -0.65 UMA: -0.60	AAF: 0.72 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.49 SAK: 0.51 UMA: 0.60	AAF: 0.39 ALF: 0.33 AGS: 0.14 DAH: 0.22 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.34 ALF: 0.68 AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
4	0.9841771, 0.9880693, 0.9965154, 0.9904878, 0.999, 0.9895322, 0.9950292	-0.8965172, 0.7695836, 0.3992047, 0.3532392, -0.6460816, 0.5472614, 0.3452207, 0.7006323, -0.3383669, 0.421621, 0.1402084, 0.4460048, -0.7832472, 0.7057979, 0.2312134, 0.4866034, -0.590711, 0.544264, 0.1768126, 0.6368714, -0.7226458, 0.5551665, 0.4538412, 0.4140419, -0.6547262, 0.6460972, 0.2081288, 0.5428632	525.8587	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.89 ALF: -0.59 AGS: -0.31 DAH: -0.72 HNW: -0.53 SAK: -0.65 UMA: -0.60	AAF: 0.72 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.49 SAK: 0.51 UMA: 0.60	AAF: 0.40 ALF: 0.33 AGS: 0.14 DAH: 0.22 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.34 ALF: 0.68 AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
5	0.9818318, 0.9882449,	-0.8965172, 0.7695836, 0.3992047, 0.3532392, -0.6460816, 0.5472614, 0.3452207,	525.8587	AAF: 0.980 ALF: 0.990	AAF: -0.89 ALF: -0.59	AAF: 0.72 ALF: 0.51	AAF: 0.39 ALF: 0.33	AAF: 0.34 ALF: 0.68

	0.9965498, 0.9911239, 0.999, 0.9890333, 0.9949012	0.7006323, -0.3383669, 0.421621, 0.1402084, 0.4460048, -0.7832472, 0.7057979, 0.2312134, 0.4866034, -0.590711, 0.544264, 0.1768126, 0.6368714, -0.7226458, 0.5551665, 0.4538412, 0.4140419, -0.6547262, 0.6460972, 0.2081288, 0.5428632		AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AGS: -0.31 DAH: -0.72 HNW: -0.53 SAK: -0.65 UMA: -0.60	AGS: 0.40 DAH: 0.66 HNW: 0.49 SAK: 0.51 UMA: 0.60	AGS: 0.14 DAH: 0.22 HNW: 0.17 SAK: 0.41 UMA: 0.20	AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
6	0.9882387	-0.8909867, 0.7537725, 0.397376, 0.3439293	525.8595	AAF: 0.980 ALF: 0.990 AGS: 0.996 DAH: 0.990 HNW:0.999 SAK:0.990 UMA: 0.995	AAF: -0.90 ALF: -0.60 AGS: -0.31 DAH: -0.73 HNW: -0.54 SAK: -0.66 UMA: -0.61	AAF: 0.72 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.50 SAK: 0.51 UMA: 0.61	AAF: 0.40 ALF: 0.32 AGS: 0.144 DAH: 0.23 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.33 ALF: 0.68 AGS: 0.41 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
7	0.95	-0.6027622, 0.4967101, 0.2858398, 0.6120133	525.8622	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.89 ALF: -0.58 AGS: -0.29 DAH: -0.72 HNW: -0.51 SAK: -0.64 UMA: -0.58	AAF: 0.73 ALF: 0.52 AGS: 0.40 DAH: 0.68 HNW: 0.48 SAK: 0.53 UMA: 0.61	AAF: 0.39 ALF: 0.33 AGS: 0.15 DAH: 0.23 HNW: 0.21 SAK: 0.40 UMA: 0.22	AAF: 0.35 ALF: 0.69 AGS: 0.41 DAH: 0.47 HNW: 0.60 SAK: 0.44 UMA: 0.51
8	0.95	0.2	525.8638	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.90 ALF: -0.60 AGS: -0.32 DAH: -0.73 HNW: -0.54 SAK: -0.66 UMA: -0.61	AAF: 0.72 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.49 SAK: 0.51 UMA: 0.60	AAF: 0.50 ALF: 0.33 AGS: 0.14 DAH: 0.22 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.34 ALF: 0.68 AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.51
9	0.95	-0.993131, 0.7850118, 0.4394743, 0.3834882	525.845	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.87 ALF: -0.63 AGS: -0.33 DAH: -0.77 HNW: -0.57 SAK: -0.71 UMA: -0.65	AAF: 0.73 ALF: 0.56 AGS: 0.44 DAH: 0.68 HNW: 0.53 SAK: 0.55 UMA: 0.67	AAF: 0.39 ALF: 0.36 AGS: 0.17 DAH: 0.29 HNW: 0.21 SAK: 0.47 UMA: 0.22	AAF: 0.42 ALF: 0.83 AGS: 0.51 DAH: 0.56 HNW: 0.74 SAK: 0.54 UMA: 0.66
10	0.95	-0.8909867, 0.7537725, 0.397376, 0.3439293	525.8586	AAF: 0.980 ALF: 0.990 AGS: 0.997 DAH: 0.990 HNW: 0.999 SAK: 0.990 UMA: 0.990	AAF: -0.88 ALF: -0.57 AGS: -0.29 DAH: -0.71 HNW: -0.51 SAK: -0.64 UMA: -0.58	AAF: 0.73 ALF: 0.51 AGS: 0.40 DAH: 0.66 HNW: 0.50 SAK: 0.51 UMA: 0.61	AAF: 0.38 ALF: 0.33 AGS: 0.14 DAH: 0.23 HNW: 0.17 SAK: 0.41 UMA: 0.20	AAF: 0.34 ALF: 0.69 AGS: 0.42 DAH: 0.47 HNW: 0.62 SAK: 0.43 UMA: 0.52

BIBLIOGRAPHY

1. Fuller WA. Measurement error models. New York: Wiley; 1987.
2. Fleiss JL, Levin BA, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, N.J.: J. Wiley; 2003.
3. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20(1):37 - 46.
4. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. Arch Gen Psychiatry 1985;42(7):725-8.
5. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-74.
6. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70(4):213-20.
7. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423-9.
8. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43(6):551-8.
9. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43(6):543-9.
10. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin 1987;101(1):140-146.
11. Cerretti DP, Lyman SD, Kozlosky CJ, Copeland NG, Gilbert DJ, Jenkins NA, et al. The genes encoding the eph-related receptor tyrosine kinase ligands LERK-1 (EPLG1, Epl1), LERK-3 (EPLG3, Epl3), and LERK-4 (EPLG4, Epl4) are clustered on human chromosome 1 and mouse chromosome 3. Genomics 1996;33(2):277-82.
12. Guggenmoos-Holzmann I. How reliable are chance-corrected measures of agreement? Stat Med 1993;12(23):2191-205.
13. Lantz CA, Nebenzahl E. Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. J Clin Epidemiol 1996;49(4):431-4.
14. Uebersax JS. Modeling approaches for the analysis of observer agreement. Invest Radiol 1992;27(9):738-43.
15. Skrondal A, Rabe-Hesketh S. Latent variable modelling. Stat Methods Med Res 2008;17(1):3-4.
16. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56-62.
17. Bollen KA. Structural equations with latent variables. New York: Wiley; 1989.

18. Loehlin JC. Latent variable models : an introduction to factor, path, and structural analysis. Hillsdale, N.J.: L. Erlbaum Associates; 1987.
19. Tinsley HEA, Brown SD. Handbook of applied multivariate statistics and mathematical modeling. San Diego: Academic Press; 2000.
20. Agresti A. Analysis of ordinal categorical data. 2nd ed. Hoboken, N.J.: Wiley; 2010.
21. Norusis MJ. Ordinal regression. In: SPSS I, editor. IBM SPSS Statistics 19 Advanced Statistical Procedures Companion: Pearson; 2011.
22. Worm base website, http://www.wormbase.org, release WS240; 2013
23. Hu PJ. Dauer. WormBook 2007:1-19.
24. Brenner S. The genetics of Caenorhabditis elegans. Genetics 1974;77(1):71-94.
25. Golden JW, Riddle DL. A pheromone-induced developmental switch in Caenorhabditis elegans: Temperature-sensitive mutants reveal a wild-type temperature-dependent process. Proc Natl Acad Sci U S A 1984;81(3):819-23.
26. Liu ZC, Ambros V. Heterochronic genes control the stage-specific initiation and expression of the dauer larva developmental program in Caenorhabditis elegans. Genes Dev 1989;3(12B):2039-49.
27. Bargmann CI, Horvitz HR. Control of larval development by chemosensory neurons in Caenorhabditis elegans. Science 1991;251(4998):1243-6.
28. Kimura KD, Tissenbaum HA, Liu Y, Ruvkun G. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. Science 1997;277(5328):942-6.
29. Kenyon C, Chang J, Gensch E, Rudner A, Tabtiang R. A C. elegans mutant that lives twice as long as wild type. Nature 1993;366(6454):461-4.
30. Ogg S, Paradis S, Gottlieb S, Patterson GI, Lee L, Tissenbaum HA, et al. The fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. Nature 1997;389(6654):994-9.
31. Paradis S, Ruvkun G. Caenorhabditis elegans Akt/PKB transduces insulin receptor-like signals from AGE-1 PI3 kinase to the DAF-16 transcription factor. Genes Dev 1998;12(16):2488-98.
32. Morris JZ, Tissenbaum HA, Ruvkun G. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in Caenorhabditis elegans. Nature 1996;382(6591):536-9.
33. Hu PJ, Xu J, Ruvkun G. Two membrane-associated tyrosine phosphatase homologs potentiate C. elegans AKT-1/PKB signaling. PLoS Genet 2006;2(7):e99.
34. Inoue T, Thomas JH. Suppressors of transforming growth factor-beta pathway mutants in the Caenorhabditis elegans dauer formation pathway. Genetics 2000;156(3):1035-46.
35. Beginners guide to OpenMx website, www.vipbg.vcu/~hmaes/OpenMx/html/NewBeginnersGuide.html#categorical-data, 2013.