INTERNATIONAL JOURNAL OF COMPARATIVE PSYCHOLOGY

2018, 31
Charles Abramson
Special Issue Editor
*Peer-reviewed**

Ordinal Pattern Analysis in Comparative Psychology: A Flexible Alternative to Null Hypothesis Significance Testing Using an Observation Oriented Modeling Paradigm

David Philip Arthur Craig and Charles I. Abramson

Oklahoma State University, U.S.A.

The data of comparative psychology generally differ from the majority of data collected within mainstream psychology in several key respects – most notably in the diversity of forms of measurement and fewer number of subjects. We believe null hypothesis significance testing may not be the most appropriate method of analysis for comparative psychology for these reasons. Comparative psychology has a rich history of performing several analyses on a few subjects due to a philosophical interest in individual subject behavior, along with group assessments. Since first being published in 2011, Observation Oriented Modeling has successfully been used to analyze individual subjects' responses from honey bees, horses, humans, and rattlesnakes. Observation Oriented Modeling is highly flexible and has allowed comparative researchers to perform a variety of assessments comparable to null hypothesis significance testing's *T*-Tests, One-way ANOVA, and Repeated-Measures ANOVA while producing easily-interpretable and, most importantly, relevant results. This paper describes the diverse manners in which comparative psychologists can assess individual and group performances without concerns of statistical assumptions and limitations that complicate assessments when employing Null Hypothesis Significance Testing.

Comparative psychology is insulated from the majority of psychology sub-fields due to inherent differences in data collection between human and animal models. In practice, measuring animal behavior is based in a fundamentally different methodology compared to measuring human behavior. Comparative data are generally collected in-line with the traditional (Michell, 2014) definition of measurement (i.e., $a=r*b$ where r is a unit and b is a magnitude). In contrast, modern human behavioral data are generally collected following Stevens' (1946) redefinition of measurement (i.e., the assignment of numbers to properties following a rule). These different definitions of measurement solidified a major difference between psychometric and physics-based science (Humphry, 2013), and comparative psychology began using both forms of measurement due to strong influences from both biology and psychology. While psychology's influence from biology and the physics-based sciences initially impressed a use of both continuous and discrete measurement, throughout the later part of the 20th century, influences from psychometric and cognitive research introduced more qualitative measures into comparative psychology (Whissell, Abramson, & Barber, 2013). The common use of these different forms of measurement across the sub-fields requires comparative psychologists use both psychometric and physics-based methods.

Measurement Redefined

After Stevens' (1946) redefinition of measurement, psychometricians' data required clarification to contend with diminishing effect sizes. After all, imprecise measurement without clear units led to relatively

Please send correspondence to David Philip Arthur Craig, Oklahoma State University, U.S.A. (Email: dpac007@gmail.com)
Support for this research was provided by grants NSF-REU (2016-1560389), and NSF-OISE (2015-1545803).
* Dr. H. M. Hill acted as action editor on this paper for the special issue. https://doi.org/10.46867/ijcp.2018.31.01.10

unclear data (John, Loewenstein, & Prelec, 2012), and these data required enhanced new data analyses methods to observe significant differences. The emphasis on operationalism and representationalism of Stevens' (1946) measurement created an artificial focus on instrumentation, application, and permissibility that drew attention away from the equivocation concerns around the actual term of measurement. Before Stevens (1946) redefined measurement in psychology, Ferguson et al. (1940) specifically referenced the measurement reported in Stevens and Davis (1938) as an example of why true quantitative measurement was impossible for psychometricians. The invention and adoption of Stevens' measurement validated and encouraged estimates of sensory events, and the focus on the four scales of measurement raised more questions about scale transformations and permissibility rather than if the proxy scale, and actual attribute being estimated, was actually a quantitative measure (Michell, 2008). The quantitative imperative (viz., knowledge and science require quantitative measurement) was left unaddressed, and the relevancy of the quantitative hypothesis (viz., measured mental attributes are actually quantitative) was largely untested. Instead, psychologists successfully emulated other sciences in competition for grant money by using quantitative methods for measures on interval and ratio scales (Solovey, 2004).

By using quantitative methods, psychometricians appeared to be performing quantitative science, and successfully avoided criticism for not adhering to the quantitative imperative (Michell, 2008). Of these quantitative methods, parametric Null Hypothesis Significance Testing, or the process of predicting there is no mean difference in populations, became the most popular analysis tool for psychometric science. Parametric Null Hypothesis Significance Testing further strengthened the psychometricians' confidence in satisfying the requirements of the quantitative imperative, and thus helped perpetuate the claims that psychology was a quantitative science. While psychometric data were treated as if they are quantitative, neither the observed measure, nor the operationalized construct, had been demonstrated to actually be quantitative. While not actually measuring quantitative data, the use of parametric Null Hypothesis Significance Testing treated the data (and most importantly, the construct) as if they were quantitative, but this quantitative hypothesis has largely been untested within psychology (Michell, 2008). Rather than address this fundamental quantitative assumption, psychometricians increasingly relied on larger sample sizes, group designs, and corrective transformations to assess their data and obtain significant, publishable results using parametric Null Hypothesis Significance Testing. Significance replaced meaningfulness, and psychologists began estimating population parameters under a hypothesis no mean differences exist in the population (Gigerenzer, 2004).

Despite this shift in measurement and the accompanying analyses, comparative psychologists and radical behaviorists largely continued measuring and collecting data in the traditional sense and conformed to Hölder's axioms of quantity (Michell & Ernst, 1997). Common comparative measures like inter-response time clearly fit Hölder's axioms (viz., communitive, associative, and transitive properties, and a sums of magnitudes are greater than individual magnitudes). Comparative psychological data also largely conformed to Hölder's additional requirements of density to be able to consider a quantitative measure as continuous (i.e., consistency of magnitudes and upper/lower bounds). Measures such as response time contained clear units, were continuous, and thus suitable for undergoing the additive multiplicative operations required when calculating a mean. These comparative data were tangible and real, and the effects of manipulations could easily be deciphered by, for example, visually displaying a cumulative number of responses across time. Comparative data were already quantitative and satisfied the quantitative imperative, so unlike the psychometricians, comparative researchers did not need to use parametric Null Hypothesis Significance Testing to validate reliance on qualitative measurement.

Null Hypothesis Significance Testing in Comparative Psychology

However, most comparative psychologists are now trained as general psychologists; given the diminishing number of course and curriculum options for students interested in comparative psychology, few other data analysis options are afforded to comparative psychology students (Abramson, 2015). Perhaps because of these shared curriculums, Null Hypothesis Significance Testing has slowly been adopted by the comparative sub-field rather than being leveraged for methodological or metaphysical reasons. In practice, the use of Null Hypothesis Significance Testing may also have been stimulated by comparative psychology's tendency to publish in a wide variety of journals, and the differences in these other sub-fields' journal standards. As comparative psychology differs in fundamental ways compared to the rest of psychology's sub-fields, and data analyses methods should be selected based on the type of data to be analyzed, the use of Null Hypothesis Significance Testing (NHST) may not be the most appropriate of available options for the comparative psychologist. Comparative researchers should not feel obliged to use NHST simply because their colleagues do so with their human models. In light of the differences between human and animal data, we pose the use of NHST may not have been the most suitable analysis method for comparative psychology. Three key aspects of comparative research encourage critically evaluating the continued use of NHST in comparative psychology:

1) Behavioral or physiological data are the only available dependent variables for comparative psychologists; after all, communication barriers between species make self-report assessments useless. Comparative psychologists have continued to explore and refine measures that are more likely to be based and occur in reality. The measurement of dependent variables based in space-time (e.g., length, duration, responses per minute, inter-response time, temperature) requires the use of units (i.e., common standards), ratios of magnitudes (i.e., consistent standards), and continuity (i.e., additive and density mathematical properties), and thus continues the traditional form of measurement. Indeed, Herman von Helmholtz (1887) was among the first to question how to know if an attribute was additive; even though Helmholtz was measuring reaction time (a continuous quantitative measure), this form of critical inquiry and assessment-checking had a firm tradition in comparative work prior to the sub-field's genesis.

In contrast, modern human psychological investigators favor both Stevens' (1946) redefinition of measurement and invented four forms of measurement: nominal, ordinal, interval, and ratio measures. Despite his clear arguments as to why ordinal data could not undergo multiplicative operations (e.g., to calculate a mean) due to a lack of continuity, Stevens' (1946) sent modern psychology toward a discussion of statistical permissibility (Velleman, & Wilkinson, 1993). Despite the rigorous methods and discussion (Schorske, 1997), it seems the majority of modern human psychological researchers incorrectly classify their data. For example, Likert data, data which can only be considered ordinal, are treated as interval data that may then be averaged. Hölder's axioms are still the benchmark to allow additive and multiplicative mathematical properties for certain types of numbers and data; psychometric data are no exception. However, the influence of psychometrics on comparative psychology is undeniable (Whissell, et al., 2013), so because of its history of measurement, comparative psychology will likely continue to collect both continuous and discrete quantitative data (along with qualitative data as psychometrics is further adopted into animal research). A method of data analysis that contends with all three forms of data (continuous quantitative, discrete quantitative, qualitative) would best serve a sub-field that collects a diverse form of data due to a mix of influence of traditional measurement, and the psychometrican's form of measurement.

2) Comparative psychological research seldom relies on introductory psychology students to serve as subjects to quickly build large Ns; animal husbandry requirements, barriers to field data collection, and other

logistical difficulties of animal data collection result in a sub-field that publishes with a relatively lower *N* compared to the standards of other psychological sub-fields. Throughout its history, comparative psychology has demonstrated a keen focus on individual subject behaviors (Mace & Kratochwill, 1986). For example, Skinner Boxes originally showed cumulative responding for individual subjects, and Skinner's qualitative cumulative curves forced the use of response rate as a dependent variable. Experiments with individual assessments of thousands of trials for a single subject clearly exhibits an interest in the individual subject (e.g., Ferster & Skinner, 1957). After all, learning occurs in an individual subject, not in a group average. Within-subjects designs were popular in radical behaviorism, and along with group analyses, single-subject and other designs with lower *N*s are still popular within comparative psychology. For example, recent subject sizes in three of the most popular comparative and social psychology journals highlight this stark contrast (Table 1).

Table 1

Comparison of Ns in Popular Animal and Human Psychology Journals

Journal	Date of Publication	Median N per Experiment	Median N per Publication
Journal of Comparative Psychology	Nov/2017	8	12
Behavioural Processes	Dec/2017	16	16
Journal of Experimental Analysis of Behavior	Nov/2017	9	8
Social Behavior and Personality	Oct/2017	298	305
Journal of Experimental Social Psychology	Nov/2017	154	810
The Journal of Social Psychology	Nov/2017	187	277

Ultimately, these lower *N*s impact the comparative psychologist's ability to generalize findings and make convincing conclusions about a population based on a relatively small sample. For example, how could comparative researchers with a sample size of four turtle doves (Lejeune & Richelle, 1982) or eight Aldabra giant tortoises (Spiezio, Leonardi, & Regaiolli, 2017) meaningfully generalize findings to a population that is not critically endangered? The use of Null Hypothesis Significance Testing (NHST) may not be the most ideal form of analysis for comparative psychology because truly representative samples are rarely assessed in the field. Data analysis methods that do not abstract to assessments of a population parameter and can perform individual subject analyses may better serve a researcher with fewer subjects. However, this inability to generalize to a population is not an inherent issue with data collection or analysis, just in the ability to draw generalizable conclusions.

3) Due to the difference in subject size and the established interest in individual subject performance, we believe comparative psychology's data analysis methods should be focused more on individual subjects and observations rather than abstracting observed data to a population parameter based on aggregate assessments. Group assessments and aggregate analyses are certainly helpful for initial analyses, but individual analyses allow researchers to further investigate the observed data, and may reveal artifacts associated with aggregate based analyses. Indeed, comparative psychology's literature is rich with well-articulated points

(Branch & Gollub, 1974; Bakan, 1967; Meehl, 1978; Skinner, 1956; Dews, 1978; Schneider, 1969; Zeiler & Powell, 1994) against dependence on aggregate, and especially mean and variance based, analyses. These cautionings should continue to encourage comparative psychologists to consider other data analysis methods that assess individual subject performances. With lower *N*s, comparative psychologists can dedicate more attention to individual subjects and rely more on within-subjects and single-subjects designs. Data analyses that favor individual observations within and between subjects, do not need data transformations, and offer standardized and clear output could be expected to especially appeal to comparative psychologists.

Alternatives to Null Hypothesis Significance Testing

Statisticians such as Russell, Campbell, Nagel, and Stevens invented a series of regression analyses that are easy to misunderstand (e.g., the focus on the p-value above an effect size, the definition of a p-value, the implications of statistical significance, meaningfulness) and misinterpret (Cohen, 1994; Gigerenzer, 2004). The statistical misuse outside of comparative psychology is likely to blame for the unprecedented replication failures experienced within the more popular sub-fields like cognitive or social psychology (Doyen, Klein, Pichon, & Cleeremans, 2012; Nosek et al., 2015). However, by remaining based in traditional measurement and demonstrating a historical interest in individual subjects, comparative psychology does not need to depend on NHST like the psychometric or cognitivist sub-fields.

More recently, NHST advocates are being challenged in many scientific fields (Woolston, 2015; Beghetto, 2014; Gigerenzer, 2004), but NHST has been criticized throughout comparative psychology's history (Skinner, 1972) likely because behavioral data frequently do not meet the assumptions required to perform NHST (Laurent & Lejeune, 1985). A field using real idemnotic measurement, with a focus on individual subjects that utilizes small sample sizes with non-normal and non-homogeneous data could benefit from exploring alternatives to NHST. Comparative psychology's continued use of NHST should be questioned, not just because of NHST's own limitations, but because there was never a need for NHST in a sub-field that was focused on individual subjects and performed real measurement.

As comparative psychology becomes a rarer course in collegiate catalog offerings (Abramson, 2015), a bridge between physics-based science and mainstream psychology diminishes, and animal researchers will continue to use the only option that is seemingly available for psychological data analysts: NHST. Bayesian methods could be classified as the top alternative option to NHST, but Bayesian methods may be hard to adopt in sub-fields that collect and report data as frequencies rather than probabilities (Gigerenzer & Marewski, 2015). Frequency-based assessments may better serve the comparative psychologist, so a different alternative to Bayesian methods may be more appropriate for animal researchers. We believe implementing frequency-based data analysis methods that focus on individual subjects may be integral to the continued growth of comparative psychology. Moreover, Bayesian methods also abstract to population parameters and thus may be less useful for a sub-field with lower sample sizes and an interest in explanatory inferences. The established focus on individuals in comparative psychology stems from a desire to describe the causal mechanisms (viz., contingencies) impacting the individual's behavior, so this established and careful use of abductive inferences may further decrease the likelihood of adoption of an 'inference machine' like Bayesian methods. Like Bayesian methods, we also do not consider Latent Class Modeling an ideal approach for comparative psychologists. While Latent Class Modeling has relatively fewer assumptions than parametric NHST, Latent Class Modeling is still designed to estimate population parameters of unobservable mental constructs. Like all Structural Equation Modeling, Latent Class Modeling is generally less helpful for comparative researchers that

perform behavioral and physiological measurement and adopt a behaviorist tendency to avoid abstraction to unobservable attributes.

Observational Oriented Modeling

Observation Oriented Modeling (Grice, 2011) is an established method of data analysis that is capable of assessing the types of data that are collected by comparative researchers, and doing so without taking an aggregate to summarize the individual observations. Fundamentally, Observation Oriented Modeling (OOM) focuses scientific descriptions on the actual observed data: individual subjects making individual responses. Philosophically, OOM empowers comparative psychology to align all of its methods with its metaphysics. The use of NHST stands in contrast with the rich tradition of quantitative measurement of individual subject behavior within comparative psychology. Practically, OOM empowers comparative psychology to actually ask questions that matter to the sub-field. NHST cannot answer the simple question: *were more responses observed in A versus B?*. NHST only allows researchers to reject a null hypothesis that the population parameters of the observed data differ when compared to a supposedly normal distribution. Prior to ever being used to analyze comparative psychological data, Grice (2011) specifically identified behaviorists (and by extension, comparative researchers) would most likely be the first psychologists to understand the benefits of OOM. Below, we explain why the flexibility of OOM's Ordinal Pattern Analysis should appeal to comparative psychologists that are interested in performing traditional measurement and assessing individual subject behavior.

Ordinal Pattern Analysis

Within the philosophy of OOM, the Ordinal Pattern Analysis (Grice, Craig, & Abramson, 2015), seems especially useful for comparative psychologists. Ordinal Pattern Analysis (OPA) is a highly-flexible assessment that allows analysts to make any number of *a priori* ordinal predictions (e.g., an FR 2 schedule of reinforcement would produce lower responding than an FR 4). The observed data are then compared to the ordinal prediction to calculate a Percent Correct Classification Value (PCC Value) before the observed data are randomized a set number of times. Each of these randomizations is then compared to the ordinal prediction to create a distribution of PCC values that are then compared to the observed data's PCC Value. The comparison of the randomizations' PCC values versus the observed data's PCC value produces a probability statistic, a chance value (*c*-value). The PCC value is an effect size and displays the number of observed data that match the ordinal prediction compared to the total number of performed ordinal comparisons. The assessment is simple at its core and thus easy to understand and leverage to draw clear conclusions. Additionally, researchers may also include the presentation of the lowest and highest PCC values obtained from the randomization comparisons along with the PCC value and *c*-value. In doing so, OPA can more easily be understood by those based in an NHST paradigm; the PCC value is an effect size, the c-value is a probability statistic, and the randomization range is analogous to a confidence interval.

When performing pair-wise assessments, OPA calculates a PCC value and accompanying c-value. However, ordinal predictions with more than two orders can also be made and assessed in OOM. When performing omnibus assessments, OOM performs both pair-wise and omnibus assessments and calculates a PCC value for all of the pair-wise assessments, and a Complete Percent Correct Classification Value (CPCC Value) for the omnibus assessment. The CPCC value compares all of the observations to the entire ordinal prediction whereas the pair-wise assessments compare all observations against segments of the ordinal

prediction. If a single subject had three observations of movement (15 cm, 12 cm, 18 cm) compared to a monotonically increasing ordinal prediction, the CPCC value would be 0 because a perfect match was not observed, but the PCC value would be 66.66%. With this CPCC value, OPA can be extended to compare any number of groups or conditions.

Moreover, OPA can perform both dependent and independent analyses. Dependent Ordinal Analyses simply compare each observation as relevant without involving other observations. For example, a two-subject pre-test versus post-test assessment would perform two ordinal comparisons (i.e., A1 vs B1; A2 vs B2). Independent Ordinal Analyses can be performed by comparing combinations of observations. For example, a two subject assessment would perform four ordinal comparisons (i.e., A vs C; A vs D; B vs C; B vs D). If comparing more than two conditions, repeated measures assessments can be performed for dependent data. For example, a two-subject pre-test versus post-test versus follow-up assessment would perform eight assessments (i.e., 1A vs 1B; 1A vs 1C; 1B vs 1C; 1A vs 1B vs 1C; 2A vs 2B; 2A vs 2C; 2B vs 2C; 2A vs 2B vs 2C). If comparing more than two groups, multiple combination assessments can be performed as a natural extension of a two group comparison. For example, a three group comparison with two-subjects each would perform 16 pair-wise assessments and 8 omnibus assessments.

Inherent Flexibility of OPA

The diversity of OPA is of particular importance for a sub-field that must contend with instrumental (viz., procedural) differences when making indirect comparisons between species. Being able to leverage similar analyses for individual subjects addresses analytical and procedural differences to better facilitate making inter-species comparisons. Rather than using different types of tests across different forms of data, OPA is just one type of test. This robustness of OPA is inherent in its methods of analysis: the ordinal comparisons, the calculated PCC value, and the randomizations.

1. A major contributor to this flexibility are the ordinal comparisons themselves; the observed data are treated in a non-parametric manner. While seemingly conservative, treating observations as discrete quantities, or as the observations are actually collected, eschews the difference between parametric and non-parametric data assessments (allowing for an even greater number of comparisons), and thus aligns the analysis with the observation. In the process, an ordinal analysis side-steps theoretical implications of violations of continuity in measurement. For example, should a clearly discrete binomial observation, such as a response (Michell 1994; 1997), be treated as a continuous observation if divided by an interval of time? Does the division of time and transformation to a response rate suddenly justify the use of parametric assessments for data that are otherwise clearly discrete? If a numerator does not adhere to the density property, division by a continuous variable may not be appropriate. OPA's treatment of the observations in an ordinal manner eschews deviating definitions of measurement (and assumptions of continuity) to allow a wide variety of forms of data to be assessable via the same analysis method. OPA may be valuable for comparative psychologists that otherwise may face difficult decisions about the composition and constitution of an observation, and thus the most appropriate NHST assessment to perform. Theoretically, OPA maintains how the observations and data are treated by the researcher; discrete and continuous data can be analyzed in the same manner. Practically, OPA bypasses assumptions requiring considerations of statistical permissibility along with qualifying and inconsistent corrections processes.

2. PCC values are an ideal effect size because the value occurs on a 0-100 scale (percent), so indirect comparisons via outcomes from a data analysis tool is not methodologically offensive when using OOM (as long as procedural considerations are made). Because OPA's PCC value is calculated purely on comparisons

of the individual observations and without the use of aggregates from a particular set of observations, the PCC values are easy to compare across datasets. As the PCC value is not calculated with an aggregate such as a standard deviation, the PCC value is not vulnerable to aggregate artifacts. Moreover, conventions of interpretation, such as for Pearson's *r*, Cohen's *d*, and η^2 are not required because a PCC is displayed as a standard percent. In contrast, many common effect sizes in NHST such as Cohen's *d* and η^2, do not have clear and absolute limits, meaning, or methods of interpretation without arbitrary guidelines. Several common effect sizes in NHST make indirect comparisons via data analysis impossible because these values are relative to the observed data set. The PCC value's absolute standard facilitates making indirect comparisons and may be a more effective effect size for comparative psychologists using common effect sizes with NHST. The PCC value is highly transparent, simple to compute, easy to understand, and thus an ideal effect size to interpret. For example, if a comparative researcher were to predict subject lever-press responding decreases across 5 trials in an extinction contingency, and the PCC value for dolphin responding was 75.00, but the PCC value for shark responding was 50.00, the PCC value alone could be used to indirectly compare extinction across species and allow the researcher to conclude dolphin responding was more sensitive to the extinction contingency than shark responding. The same type of indirect comparison of η^2 in a repeated measures ANOVA would not be as easy to interpret.

3. Randomization-generated probability statistics are more robust than assessments with requirements that the data (and assumed population) occur on a specific type of distribution (Manly, 2006). By generating a randomized distribution, observations are catered to what was actually observed, not what is assumed could potentially occur for some population. Concerns around kurtosis, skew, and assumptions of normalcy are irrelevant with randomization. Unlike the *p*-value, the *c*-value is transparent and easy to interpret and use to perform abduction because it simply shows the number of randomized assessments that better fit the ordinal prediction compared to the observed data.

Extended Flexibilities of OPA

While OPA compares individual observations, the assessment is not limited to assessing individual subjects' responding. Group assessments can be performed by pooling individual subjects' data and comparing every groups' individuals' responses. Pooling simply combines observations, but unlike an aggregate, no attempts to describe the individual data are attempted when pooling observations. Instead, OOM compares the observed data, not descriptions (such as measures of central tendency) of the data. As statistical power is irrelevant with OOM, performing both group and individual subject analyzes is permissible and encouraged. Moreover, by using pooling to perform dependent analyses, OOM allows researchers to perform assessments with unequal response observations. This adaptability is greatly beneficial to comparative psychologists that perform measurement in the field and cannot perfectly control variables such as trial/session duration. This flexibility may also appeal to researchers attempting to make indirect comparisons across species. Statistical limitations of data analyses that make no real theoretical violations, such as unequal numbers of observations or insufficient degrees of freedom, introduce artificial barriers that are not based in the organic reality of the observations, or the context in which the data are collected. Fields wherein data collection cannot occur in rigid circumstances may be especially stifled by the adoption of assumption-laden data analysis methods.

Indeed, OPA is extremely flexible in part because it eschews many of the assumptions traditionally expected in data analysis due to the monopolistic omnipresence of NHST. Dinges et al. (2013) outline four clear differences in assumptions between OOM vs NHST.

1. OPA does not make assumptions regarding *continuity*. While parametric NHST's assumption of continuity is more frequently violated by psychometric measurement, comparative psychology does use dependent variables with discrete quantities (e.g., response counts) that should not undergo multiplicative properties like means (or medians with equal *N*s). The frequent violation of the continuity assumption is troubling because there is no corrective action to perform when Hölder's axioms are not met. When parametric NHST is used to assess qualitative data, the actual transformations of the data violate the mathematical properties of the observations (e.g., additive and multiplicative). Calculating a mean for a Likert scale, or any qualitative data, is not mathematically permissible as fundamental properties about the observations have not been met, and many of these aggregates (e.g., standard deviation, mean) that are critical to calculate a probability statistic are functionally meaningless because these aggregates cannot possibly exist in reality. The eschewing of the continuity assumption makes the Ordinal Analysis versatile and capable of analyzing all forms of quantitative data. Measures of both time (continuous) and responses (discrete) can be observed with the same test, using the same calcuations, and providing the same values (PCC and *c*-value). Treating data as discrete quantities aligns with how observations are generally recorded (viz., counting); continuous data are rarely recorded continuously. By treating observations in an ordinal manner, frequencies can be compared in the same manner as numerical relations. OPA does not assume the collected data have been measured, in the strictest sense, and thus allows for qualitative and quantitative data analyses.

2. OPA does not make assumptions regarding *homogeneity* (or sphericity). While parametric NHST is dependent on calculating aggregates such as means, variances, standard deviations, and sums of squares, OOM never uses an aggregate for data analysis purposes. Comparative psychologists will likely encounter homogeneity concerns when making comparisons between species; variability in discrimination capabilities in honey bees compared to rats or pigeons shouldn't prevent direct species comparisons. It is reasonable to expect species' responding will differ in different ways, and comparative psychologists need a data analysis tool that allows direct comparisons of these types of data.

3. OPA does not compare observations to a hypothetical distribution, so topics associated with *normality* are all eschewed. OPA does not use alpha-levels, and thus does not have concerns around power or degrees of freedom. By using randomizations to generate distributions, OPA empowers researchers to perform data analyses on smaller subject sizes. Given comparative psychology generally assesses smaller sample sizes compared other sub-fields within psychology, this feature of OPA is especially relevant for animal researchers. However, the gains of side-stepping assumptions around normality allows researchers to analyze the same data in numerous ways without artificial limitations surrounding the number of allotted assessments that can be performed on a series of data. For this reason, OPA allows comparing the observed data to any number of ordinal predictions to determine which ordinal prediction best characterizes the data. OPA removes concerns related to retesting previously assessed data, such as when a new species' data can be compared to historical information in the literature. Additionally, assessing different aspects of the observed data (e.g., a post-reinforcement pause versus general inter-response times) doesn't lead to conceptual issues related to power or degrees of freedom.

4. OOM does not make assumptions regarding *dependency*. OOM uses the same mathematical process to assess dependent observations as it does to assess independent observations. This means longitudinal data could be analyzed alongside cross-sectional data. Independent and dependent *T*-Tests do not use the same exact methods of analyzing independent vs dependent data. The only difference between dependent data versus independent data assessments in OPA would be to perform combination assessments for independent assessments. Unlike a dependent *t*-test, no subtraction is performed in OPA.

OOM's adaptability allows for analyses of data collected from a wide variety of procedures, and OOM has successfully been used to analyze established and common assessments performed by comparative psychologists and radical behaviorists. OOM has been utilized to assess data collected from discrimination (Dinges et al., 2013), fixed interval (Craig et al., 2014, Craig & Abramson, 2015), reversal (Abramson, Craig, Varnon, & Wells, 2015), and fixed ratio (Place, Varnon, Craig, & Abramson, 2017) procedures. Within the comparative literature, OOM has been used to assess responding in honey bees (Dinges et al., 2013, Craig et al., 2014), horses (Craig, Varnon, Pollock, & Abramson, 2015), and rattlesnakes (Place et al., 2017). Observation Oriented Modeling was not designed with animal researchers specifically in mind, but its flexibility naturally appeals towards comparative psychology.

Examples of Ordinal Pattern Analysis in Comparative Psychology

Prior to the development of OPA within the context of OOM, ordinal analyses had been recommended as a solution to effectively analyze individual data within the comparative literature (Gentry, Weiss, & Laties, 1983). Additionally, Thorngate and Carroll (1986) specifically recommended the use of an Ordinal Pattern Analysis, and have successfully published an alternative Ordinal Pattern Analysis (Thorngate & Edmonds, 2013) than the OPA within OOM that is described here. Thorngate's OPA and OOM's OPA are similar in their pairwise ordinal comparisons; Thorngate's Index of Fit is analogous to OOM's PCC value. However, there are two main differences between Thorngate's OPA and OOM's OPA. First, Thorngate and Ma (2016) indicate their OPA only performs pair-wise assessments to calculate their Index of Fit (like OOM's PCC value). However, OOM can also calculate a CPCC value to perform omnibus comparisons (e.g., A < B < C), so OOM is not limited to only performing pair-wise assessments. Second, Thorngate and Ma (2016) indicate a general disinterest in calculating a probability statistic, so their OPA does not have a process similar to OOM's randomizations and subsequent comparisons of the randomized data to the ordinal prediction. At their core, these different OPA methods are very similar. While the use of Ordinal Analyses is not a novel idea within comparative psychology and psychology in general, OOM's OPA was originally published as a method of analysis for comparative research to assess individual honey bee data. The following four honey bee investigations demonstrate OOM's OPA's flexibility across a variety of common comparative and behaviorist protocols.

OOM's OPA was first developed, described, and published to circumvent a repeated measure ANOVA's inability to properly assess the inter-session-interval (inter-visit-interval) between honey bees visiting an artificial flower and returning to the hive (Craig et al., 2012). In the experiment, Craig et al. (2012) prevented honey bees from returning to the hive for zero, five, or ten minutes by trapping the subject in the artificial flower (post-reinforcement delay). The honey bee data contained extreme outlier observations (intervals frequently ranged between five to ten minutes, but hour-long intervals were observed and terminated a subject's session), insufficient degrees of freedom (due to subject size and number of observations despite $N = 10$ per group), and 'missing' data (subjects were free-flying and could cease returning to the artificial flower). Based on the four conditions, five groups, and the goal of collecting twenty trials' worth of data for each subject, a twenty order prediction was generated under the hypothesis that subjects would fly back to the flower at increasingly shorter intervals if never trapped, but would fly back to the flower at increasingly longer intervals if trapped. Craig et al. (2012) observed groups and individuals with more extreme post-reinforcement intervals better fit the twenty order prediction, so trapping subjects for a longer duration resulted in longer inter-visit-intervals. While many individuals fit this pattern, a handful of individual subjects deviated from these general trends. Further descriptions and analyses of these data (along with a comparison between OPA and NHST) appears in Grice, et al. (2015).

Dinges et al. (2013) investigated aversive discrimination learning sex-differences in honey bees and was the first to generalize the comparative capabilities of OPA. Dinges et al. (2013) assessed aversive learning in a shuttle-box that could deliver shock on either half of the apparatus. In three separate experiments, OPA was utilized to assess behavior within trials (to assess individual learning), and between groups (to compare: master vs yoke, worker vs drone, and color discrimination of yellow versus blue). Indirect comparisons of PCC values from different ordinal assessments revealed stark individual differences within conditions and sexes. Overall, masters better matched the ordinal predictions than yokes, worker subjects tended to better match the ordinal predictions than drones, and both workers and drones tended to more accurately discriminate when shock was paired with the color blue rather than yellow. While Dinges et al. (2013) measured continuous temporal variables (e.g., response frequency, and latency to respond), several of the observations were partitioned into bins in order to perform comparisons within a single trial. The method of separating a single trial into bins is common within comparative psychology, but the practice loses the granularity of continuous responding and arguably requires discrete analysis and may not meet the assumptions of continuity. Moreover, responses are discrete observations that were then binned to create an arbitrary response rate; this transformation is frequently performed in comparative analyses, but this practice is questionable. For these reasons, the use of NHST to assess the data reported in Dinges et al. (2013) via a repeated measures ANOVA would require violations of critical assumptions. Practically, the most disruptive consideration was if the response frequency would be considered discrete or continuous. In order to demonstrate the impact of a potentially incorrect decision to treat discrete data as continuous in NHST, Dinges et al. (2013) also inappropriately analyzed the data with both Freedmen's assessments and repeated measures ANOVAs and treated the data as if they were continuous. Indeed, treating the data as being continuous resulted in significant *p*-values whereas the same data were not significantly different when considered discrete. Given the division of responses into equal bins creates constant denominators, the data are not fundamentally different, so the fact that separate NHST assessments of the same data can be used to draw different conclusions is worrisome. Moreover, due to violated sphericity assumptions, both a Greenhouse–Geisser correction and Huynh–Feldt correction were performed as a demonstration of the differences in correction processes, and Dinges et al. (2013) reported some cases wherein the latter was significant when the former was not. This inconsistency is also troubling. If a researcher were to select the wrong correction method, incorrect conclusions may be drawn, and clear guidance on the proper usage of a Greenhouse–Geisser correction versus Huynh–Feldt correction is not established. These types of complications that are more common in behavioral data can be by-passed with OPA.

Craig, Varnon, Sokolowski, Abramson, and Wells (2014) further established the generalizability of OPA by extending the number of independent group comparisons from two orders (analogous to an independent *t*-test) to three orders (analogous to a one-way ANOVA). This generalization was needed in order to compare the response rates of three different groups of honey bees' response levels during an extinction contingency. Similar to Craig et al. (2012), use of NHST by Craig et al. (2014) would have provided highly qualified results because of unequal numbers of observations associated with voluntary attrition (honey bees were free-flying). Additionally, homogeneity assumption violations, normalcy assumption violations, and an insufficient degrees of freedom would have prevented sufficient analyses of the data. Craig et al. (2014) exposed honey bee responding to fixed interval schedules of reinforcement using an artificial flower, and analyzed three of the more popular measures of temporal control (i.e., binned response rates, post-reinforcement pauses, cumulative curves). The three-order extinction ordinal prediction was: FI 0-sec < FI 15-sec < FI 30-sec. The pair-wise analyses indicated the data matched this ordinal prediction fairly well (PCC = 0.67, *c*-value = .001), but the three-way ordinal prediction was far less impressive despite the failure of all 1,000 randomizations to be better characterized by the ordinal prediction than the observed data (CPCC = 0.34,

c-value = .001). While insightful, additional pair-wise analyses were required to explain the distinction between the pair-wise ordinal comparisons versus the complete three-order comparison; a direct pair-wise comparison between the two experimental groups (FI 15-sec < FI 30-sec) revealed no clear difference in responding in an extinction contingency between these groups of subjects. Craig and Abramson (2015) performed a more detailed level of analysis of several measures of temporal control in honey bees (i.e., cumulative curves, response bin levels, quarter lives, IRTs, response duration, and trial duration). None of these measures displayed clear evidence of individual subject responding coming under temporal control, but pooling groups' individuals' responses did manifest evidence of temporal control. OPA empowers researchers to add these types of clarifying statements to group analyses; if individual subjects do not fit a researchers' model (e.g., an ordinal prediction), conclusions based solely on group analyses may be an analytical artifact.

Abramson et al. (2015) measured binomial response data (proboscis extension, or no proboscis extension) in honey bees across multiple sessions in a reversal protocol, and assessed honey bee reversal behavior under varying doses of ethanol. Given the observed data were binomial, dependent, and had multicollinearity and quasi-complete complications when comparing CS+, CS-, and US dependent variable, a logistical regression assessment was not ideal for the collected data. With a reversal design, OPA compared responding before and after the reversal and produced two orders of observations; with two orders of observations, a total of three ordinal predictions can be made (A > B, A < B, A = B). To thoroughly investigate the effect of ethanol on reversal learning, 510 ordinal comparisons were performed across three experiments using these three ordinal predictions. Given the reversal's design, many of the ordinal predictions comparing the CS+ responding or the CS- responding aligned with, or countered, the experimental hypothesis for some assessments. CS+ and CS- between- and within-group analyses were both performed, individuals were assessed on their own and were pooled into appropriate groups, and the comparisons combined to reveal a clear trend in CS+, CS-, and US responding. Three main-effects were assessed: reversal (control or experimental), ethanol dose (0%, 2.5%, 5%, 10%, and 20%), and ethanol administration (administered prior to training, or after training). Unsurprisingly, OPA revealed CS+ responding decreased and CS- responding increased after a reversal, and US responding did not differ between control subjects versus subjects undergoing a reversal. OPA also revealed higher doses of ethanol decreased CS+ and US responding, but did not seem to impact CS- responding. After observing these main-effects, Abramson et al. (2015) further solidified the generalizability of OPA by performing two notable interaction assessments. Using OPA, a reversal-dose interaction was observed; at higher doses, CS+ and CS- responding were more resistant to the reversal contingency change whereas the responding for subjects in the lower dosage groups displayed stronger evidence of reversal learning. Beyond this two-variable interaction, a reversal-dose-administration interaction was observed for CS- responding; administering higher doses of ethanol just prior to the reversal contingency produced higher CS- responding while administering the same doses of ethanol prior to the original discrimination training produced lower CS- responding. In this sense, Abramson et al. (2015) refined the PCC comparison process to approximate a factorial ANOVA analysis.

OPA's flexibility and ability to analyze a wide-variety of data was pioneered using honey bee data. Since establishing OPA analysis methods in honey bees, OPA has since been used to assess data from horses (Craig, Varnon, Pollock, & Abramson, 2015), humans (Jaeger, Cox, Craig, & Grice, 2016; Jordan & Thomas, 2017), and rattlesnakes, (Place et al., 2017). However, these investigations not only generalized the types of species data to have been assessed with OPA, but further extended the types of procedures to have been assessed with OPA. Further examples of OPA's generalizability exemplify its potential utility for comparative researchers.

Craig et al. (2015) employed OPA to describe horse responding under fixed interval and peak procedure contingencies, and observed clear trends in horse responding coming under temporal control. Craig et al. (2015) administered three baseline sessions for all subjects before switching to either an FI 60-sec, FI 90-sec, or FI 180-sec contingency, and as the final contingency, subjects were exposed to the peak procedure. Similar to the previous honey bee fixed interval assessments, Craig et al. (2015) assessed multiple measures of temporal control including response levels via bins, quarter life, latency to first response (post-reinforcement pause), break point (when responding shifts from low levels to higher levels), and inter-response time. These analyses in OPA revealed horse responding under fixed interval contingencies was better categorized as 'break-and-run' rather than 'scalloped', and longer fixed intervals produced clearer evidence of temporally controlled responding. To analyze the peak procedure data, each trial was divided in half, and an ordinal prediction was made for each half of the trial: responding was expected to monotonically increase for the first half of the trial, and monotonically decrease for the second half of the trial. All eleven subjects' responding monotonically increased throughout the first half of the trial, and seven of the eleven subjects' responding monotonically decreased throughout the second half of the trial. In addition to dividing the peak procedure trial in half, Craig et al. (2015) also assessed the entire peak procedure trials under the prediction responding would increase until the trial was half-way complete, and then decrease for the remainder of the trial. Four of the eleven subjects' responding fit the full prediction, so this indicates responding did 'peak' approximately half-way through the peak procedure trial for these subjects.

In a truly comparative analysis, Craig (2015) performed extensive indirect comparisons between honey bee and horse responding on FI scheduled of reinforcement via the usage of CPCC values and found horse responding came under stronger temporal control than honey bee responding via this method. Four bin response frequency ordinal assessments revealed horse responding produced higher CPCC values than honey bee responding on break-and-run ordinal predictions (e.g., $1 = 2 < 3 < 4$). Additionally, between-schedule comparisons of quarter life, index of curvature, and latency to first response reveled increasing the duration of a fixed interval schedule increased these measures of temporal control, and that horses produced higher PCC values of these assessments than honey bees. Moreover, individual subject analyses displayed higher diversity in performance in honey bees while horse responding was comparatively more uniform. Additionally, horses generally produced monotonically decreasing inter-response times within an interval while honey bee responding produced higher PCC values when comparing the data to a monotonically increase prediction (i.e., honey bees took longer to respond as the fixed interval continued while horses responded increasingly rapidly as the interval continued). OPA allows comparative researchers to indirectly compare species' performances across a variety of measures due to the simplicity of the PCC and CPCC values. These types of assessments in NHST would not be possible.

Place et al. (2017) collected response rate data for rattlesnake responding in varying fixed ratio schedule of reinforcement contingencies, and compared responding across conditions via the use of combinations. Rather than simply performing a three-way comparison like Craig et al. (2014), Place et al. (2017) compared data in shaping, CRF, FR1, FR2, FR3, and extinction conditions. Because of the number of conditions, Place et al. (2017) assessed all six of these combinations in a single ordinal assessment, but also performed more narrow comparisons (e.g. FR1 < FR2 < FR3, Shaping < FR1 < FR2). Place et al. (2017) also performed within-condition assessments and expected response rates would monotonically increase across shaping sessions, but monotonically decrease across extinction conditions. For the first time since OOM's inception, Place et al. (2017) deviated from the analysis of individual data, and instead assessed trial aggregates for individual subjects (i.e., rather than assessing responses via bins, the total number of responses in a session were pooled into appropriate conditions, and conditions were then compared). Additionally, the IRT analyses detailed in Place et al. (2017) assessed trial aggregates, and compared the usage of means or medians in OPA;

clear differences in OPA's results were observed when comparing median versus mean ordinal predictions during the extinction condition. This difference in aggregate-based comparisons serves as a reminder of the potential pitfalls associated with aggregate-based data analysis methods; which measure of central tendency is most appropriate when not relying on individual data analyses? While these types of questions encourage the use of individual analyses tools, Place et al. (2017) effectively used an individual data analysis tool to assess measures of central tendency in a given trial. With OPA, researchers are able to assess aggregates (e.g., trial average IRTs) as well as the observed individual data (individual IRTs with a trial), so researchers are certainly not limited to only performing individual analyses of raw data. So long as the measure is quantitative, appropriate aggregates are calculated, and the interpreter of the data analysis recognizes that artifacts and abstractions may remove the analyst from the real observations, there is nothing inherently preventing a researcher from using OPA to assess data summaries. Of course, this aggregate usage is not how the tool was intended to be used, but OPA certainly does not prevent a researcher from performing this type of assessment. Individual observations, as well as summaries of these observations, can be analyzed with OPA.

Ultimately, comparative researchers need to be able to draw comparisons between animal and human behavior. While comparative psychologists may be more cautious about the usage of self-reports, Jaeger et al. (2016) used OPA to assess muscle-twitch differences between self-reported extraverts, introverts, and ambiverts at varying decibel levels (50 dB, 60 dB, 70 dB, 80 dB, 90 dB, and 100 dB). In doing so, Jaeger et al. (2016) generalized the types of data to have been analyzed by OPA; for the first time, physiological data (rather than only behavioral) data were assessed. Response frequency, response amplitude, and muscle tension were assessed under the prediction introverts would produce higher metrics compared to extraverts, and ambiverts would produce middling metrics (i.e., introverts > ambiverts, introverts > extraverts, ambiverts > extraverts). From the pair-wise predictions, ambiverts did not produce response frequency data that differed from extravert or introvert data, but only comparing introverts against extraverts produced a far better match to the ordinal predictions. Via indirect comparisons of PCC values, Jaeger et al. (2016) also observed louder stimuli revealed starker differences between self-reported personality types, and introverts were more likely to respond at higher frequencies and amplitudes than extraverts.

Comparative psychology is frequently type-cast as animal research, but this caricature of comparative psychologists is narrow and dismisses the common interests and methods comparative psychology shares with other psychological sub-fields. For example, infant developmental psychologists must also contend with a language barrier, and thus are more inclined towards physiological data and behavioral scoring rather than self-reports. Andrew (1963) hypothesized facial expressions evolved in primates because these behaviors could increase social communication, and later research (Izard, 1994) into innate infant facial behaviors help support these comparative and evolutionary hypotheses. In a similar vein, Jordan and Thomas (2017) used OPA to analyze the duration and intensity of contagious human infant facial affect as recorded on an *AFFEX* scale (see Izard, Dougherty, & Hembree, 1983) and extended the generality of OPA to include analyses of qualitative, categorical data. The *AFFEX* scale involves observing three regions of a participant's face and then rating intensity and duration of movement on a 5-point Likert scale. Rather than employ Kruskal–Wallis and Mann–Whitney U non-parametric assessments, OPA was used to perform pairwise and omnibus ordinal comparisons between three stimulus conditions (audio, audio-visual, control) and two age groups (5 months and 10 months). Jordan and Thomas (2017) assessed changes in affect across these conditions and within subjects and found the audio and audio-visual stimulus conditions produced higher positive affect compared to the control condition, but did not observe a clear difference between the audio and audio-visual conditions; in contrast, negative affect differences were not clearly observed between conditions. Jordan and Thomas (2017) also compared responding in infants aged 5 months compared to 10 months with the expectation older infants would have higher affect responding, but did not observe a clear age difference in negative and positive affective

responding for these pair-wise ordinal assessments when pooling appropriate conditions or ages. In addition to behavioral and physiological data, OPA is a clear alternative for researchers collecting (or assigning) Likert scales as a form of qualitative data collection.

In perhaps its widest deviation from comparative psychological methods, OPA has been used to analyze historical data from the 15th and 16th centuries to understand how slave population frequencies in the Crimean Khanate varied across years based on the Ottoman Empire's economic fluctuations (Broyes, 2014). For these data, Broyes (2014) collected archival records of slave raids, and thus estimations of the number of slaves from the Crimean Khanate for a given year. To test the hypothesis that increases in economic prosperity were associated with increases of slave frequencies, Broyes (2014) used bureaucratically recorded economic data from the Ottoman Empire to create ordinal predictions in OPA. The Ottomon Empire saw general increases in economic prosperity up until approximately 1575 when Anatolia suffered great economic downturn, but saw general economic improvement thereafter and until the end of the 17th century. Broyes (2014) assessed the difference in slave frequencies before and after 1575 and found approximately 65% of the reported slave frequencies after 1575 were greater than frequencies prior to 1575, and a thousand randomizations of the collected data could not produce a single occurrence that better matched the ordinal prediction than the observed data. While these data and hypothesis may not be of specific interest to comparative researchers, this type of investigation could be helpful for population ecologists and evolutionary psychologists and serves as a stark example of the flexibility of OPA. Moreover, Broyes (2014) is an excellent example of how researchers can generate ordinal predictions based on hypotheses and established theories from a variety of fields and interests.

Discussion

The above use of OPA has allowed for comparative, human, and historical researchers to perform statistical analyses without the use of the following different NHST assessments:

- Dependent/Independent *T*-Tests / Chi-squares (Dinges et al., 2013)
- One-Way ANOVA (Place et al., 2017)
- Repeated Measures ANOVA / Freedman's (Craig et al., 2012)
- Factorial ANOVA / Logistical Regression (Abramson et al., 2015)
- Kruskal–Wallis / Mann–Whitney U (Jordan & Thomas, 2017)

The above comparison between NHST and OPA is intended for analogical purposes; in actuality, OPA offers far more to comparative researchers than simply being a mere substitute for NHST because of three major differences between the methods.

1. Individual vs Aggregate Analyses

OPA, and OOM in general, does not depend on aggregate analyses like NHST. Instead, OOM assesses the actual observations as they occurred in reality without performing any transformations, or relying on summaries of the observations. If only the aggregate behaviors are compared, researchers fail to ask a critical follow-up question: how well do the individuals compare? Moreover, analyzing the observed data as they were collected (in an individual, discrete manner) reduces the requirements and assumptions of the data to be analyzed. Functionally, this allows OPA to analyze all forms of measurement, including both traditional and the psychometrican's measurement. Nominal data frequencies, ordinal/discrete observations, as well as

discrete and continuous measures can all be analyzed in the same manner. In this sense, OPA makes no assumptions that attributes are quantitative; instead, OPA treats no attributes as being quantitative (i.e., aligns with the quantity objection). If trained under the popular measurability thesis (i.e., some attributes are measureable), siding with the quantity objection may be uncomfortable, but Michell (1999) articulates the measurability thesis is largely based in the embedded quantitative imperative and presumption that quantitative data are better than qualitative data.

OPA is conservative in that the method treats measured quantitative data (either discrete, or continuous) as ordinal data. In doing so, the assessment eschews many assumptions of parametric testing and increases its generalizability, but does render a researcher's analyses as qualitative. If entrenched in the positivism of the quantitative imperative, qualitative analyses may seem to have less merit. However, the quantitative imperative has likely contributed to much of the statistical misuse in psychology. Critically evaluating the relevancy of the perhaps antiquated quantitative imperative (Michell, 2003), especially when quantitate observations may not be continuous at a quantum and micro level (Craig, 2016), may serve comparative psychology well. OPA is more robust, makes fewer assumptions, and provides clearer output to better allow a researcher to perform abduction. Perhaps OPA's gains of flexibility outweigh the loss of fidelity in treating continuous quantitative measures as ordinal frequencies. Ultimately, OPA's focus on the individual observations rather than transformations of these observations may be more appropriate for a sub-field with a historical sensitivity towards extrapolating towards unobservable constructs, like an aggregate.

2. Sample Observations vs Population Parameters

The most critical difference between OOM and NHST is that OOM does not attempt to estimate a population parameter. This means OOM and NHST ask fundamentally different questions and provide different output. NHST asks a basic question: does a difference exist? Indeed, Fisher (1955, 1956) indicated NHST should be used for initial testing when little is known about the phenomena; NHST does not get an analyst very far because of the tendency to treat null as nil. The method of posting an alternative to the null hypothesis frequently introduces concerns about the direction of an implied difference when the null is rejected, and the alternative is never directly tested. Because one alternative is all that is frequently offered (rather than divide an alpha level), a researcher may incorrectly make an assumption about the direction of this difference when simply observing a significant *p*-value (i.e., make a type III error). OPA asks a different question: how well do the data conform to an ordinal prediction? Rather than only asking if a difference exists like NHST, OPA asks the specific direction, or directions, of observed differences.

The next logical question would then be: what is the magnitude of the difference for the actual observations as they were collected? NHST cannot answer this question without abstracting to a population parameter and without calculating an aggregate, nor can OPA if used strictly within the framework of OOM due to the need to calculate an aggregate (e.g., sum of differences). After performing analyses in OPA to identify the direction of a difference, traditional quantitative measures could then undergo subtraction to reveal the magnitude of the difference, but this would require aggregate comparisons (via additive transformations) and would counter the metaphysics of OOM by losing focus on the actual observations themselves.

Additionally, by not extrapolating to estimating a population and not assessing a null hypothesis of nil population differences, Type I and Type II errors are not relevant with OPA. There is no possibility of incorrectly rejecting, or failing to reject, a null hypothesis of nil differences in a population parameter. Hence, the only relevant error would be left to a researcher drawing a conclusion not supported by the PCC value, but

this would be a failure of abduction, not of OPA. Ultimately, OPA's focus on individual subjects rather than estimating a population parameter may be more appropriate for a sub-field with relatively small sample sizes.

3. Simplicity vs Complexity

Of greatest practical importance: OPA offers simple and transparent output that can be clearly interpreted on a standard (0-100 for the PCC value, and 0-1 for the *c*-value). Analysis tools must provide researchers with the ability to draw clear conclusions and make abductive comparisons without requiring qualifying statements due to violated assumptions, the appropriateness of a particular assessment, and a complete dependence on the aggregate rather than the object of interest – the individual. The simplicity of OOM creates its transparency and its flexibility. Rather than needing a decision tree to determine the appropriate form of assessment to use, OPA assesses a wide variety of data in the same manner.

In our experience, newer students of statistics struggle with understanding the definition of *p*-values, and the rampant implementation of NHST. This struggle is not a deficit at these newer students' learning abilities, or the learning-curve of introductory statistics courses, but speaks to the inherent confusion imbedded into NHST. NHST requires years of training because its methods are so counter-intuitive and seemingly complex. Simplicity and transparency in methods should be celebrated, but psychology leveraged overly-complex methods, appeared as a quantitative science, and has largely ignored criticism of its quantitative analyses (Michell, 2008). Ultimately, OPA's flexibility may be more appropriate for a field that assesses continuous and discrete quantities along with qualitative observations.

Conclusion

While the above examples of OPA's wide utility serve as clear evidence of why OPA is a viable alternative to NHST, the prevailing weakness of OPA, and OOM in general, is psychology's hesitancy to consider alternative data analyses compared to NHST. This criticism is an appeal to popularity, and is not a substantive weakness of the data analysis method itself. However, OPA does have a handful of identified minor weaknesses. The most obvious is the treatment of truly quantitative data as ordinal data. Because only ordinal comparisons are made, OPA can be extremely sensitive. For example, OPA does not adjust for practically identical data differences (e.g., a comparison of 598.21 versus 598.209). OPA has a threshold adjustment feature (Classification Imprecision value) to contend with this high-degree of sensitivity. Using this feature, researchers can 'pad' the observed values' analyses to be less sensitive by a fixed value (e.g., an analyst could set this threshold value to '1.00' to require an observed ordinal difference be at least '1.00' for the data to be considered to substantially different). This threshold adjustment feature has not yet been utilized in OPA due to requirements of an arbitrary threshold value being selected by the analyst, but this threshold adjustment feature can be used to reduce the sensitivity of OPA's analyses.

Computational space can be a practical limitation of the Ordinal Analysis, like all individual data analyses. However, OPA's use of combinations can result in hundreds of millions of ordinal comparisons if enough observations are collected. This is not especially impressive, but once the data are randomized thousands of times, billions of comparisons may actually be made via OPA to generate the *c*-value. In this sense, OPA may discourage the comparison of very large sample-sizes or observations in a single ordinal pattern assessment, or may just require usage of the PCC value rather than a *c*-value in these types of circumstances. Perhaps the limits of computational memory serve as a practical limitation to indicate the

researcher may be more interested in aggregate or populations than assessing individuals; for these types of researchers, OPA's *c*-value may not be the most appropriate alternative at a fundamental level (though the PCC value may still be a preferable effect size). Outside of its conservative treatment of continuous observations as orders and modern computer memory limitations, OPA's flexibility is limited to its users' creativity. As long as the user can articulate ordinal hypotheses to describe a phenomenon (or construct), OOM's Ordinal Pattern Analysis can analyze the provided available data.

Comparative psychology is positioned to be able to appreciate the need to shift use away from NHST to an option that may more suitable. Comparative psychology's established interests in individual observations, individual subjects, and the use of diverse forms of measurement does not align with other psychology sub-fields – why then, should comparative psychology's data analysis methods? The continued adoption of NHST within comparative psychology simply because these forms of assessments are used in other more mainstream sub-fields may stifle the growth of comparative psychology. Comparative psychology needs a more flexible data analysis option that allows for a wide variety of direct and indirect comparisons and produces standard and easily interpretable results. OOM's Ordinal Pattern Analysis is a viable contender as a preferred analysis tool for a sub-field with an established sensitivity to continuity and quantitative concerns, a history of observing individual subjects, a need for flexible assessments, and a suspended willingness to extrapolate to unobservable constructs such as a population parameter, or aggregate.

References

Abramson, C. I. (2015). A crisis in comparative psychology: Where have all the undergraduates gone? Additional comments. *Innovative Teaching, 4*, 7.

Abramson, C. I., Craig, D. P. A., Varnon, C. A., & Wells, H. (2015). The effect of ethanol on reversal learning in honey bees (*Apis mellifera anatolica*): Response inhibition in a social insect model. *Alcohol, 49*, 245–258.

Andrew, R. J. (1963). The origin and evolution of the calls and facial expressions of the primates. *Behavior*, *20*, 1–109.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation*. San Francisco, CA: Jossey-Bass.

Branch, M. N., & Gollub, L. R. (1974). A detailed analysis of the effects of d-amphetamine on behavior under fixed-interval schedules. *Journal of the Experimental Analysis of Behavior*, *21*, 519–539.

Broyes, S. C. (2014). *Slavery and the Ottoman-Crimean khanate connection* (Master's Thesis). Retrieved from: https://hdl.handle.net/11244/14742

Beghetto, R.A. (2014). Toward avoiding an empirical march to nowhere. *Psychology of Aesthetics, Creativity, and the Arts*, *8*, 18–20.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Craig, D. P. A (2015). *An assessment of multiple measures of honey bee (Apis mellifera ligustica) and horse (Equus ferus caballus) responding on fixed interval schedules: An individual versus aggregate analysis* (Doctoral dissertation). Retrieved from: https://hdl.handle.net/11244/33403

Craig, D. P. A. (2016). Shedding the quantitative imperative. *Social Behavior Research and Practice*, *1*, 10–12.

Craig, D. P. A., & Abramson C. I. (2015). A need for individual data analyses for assessments of temporal control: Invertebrate fixed interval performance. *International Journal of Comparative Psychology - Special Issue on Timing and Time Perception*, *28*, 1–39.

Craig, D. P. A., Grice, J. W., Varnon, C. A., Gibson, B., Sokolowski, M. B. C., & Abramson, C. I. (2012). Social reinforcement delays in free-flying honey bees (*Apis mellifera L.*). *PLoS One, 7*, e46729.

Craig, D. P. A., Varnon, C. A., Pollock, K. L., & Abramson, C. I. (2015). An assessment of horse (*Equus ferus caballus*) responding on fixed interval schedules of reinforcement: An individual analysis. *Behavioural Processes, 120*, 1–13.

Craig, D. P. A., Varnon, C. A., Sokolowski, M. B. C., Abramson, C. I., & Wells, H. (2014). An assessment of fixed interval timing in free-flying honey bees (*Apis mellifera ligustica*): An analysis of individual performance. *PLoS One*, *9*, e101262.

Dews, P. B. (1978). Studies on responding under fixed-interval schedules of reinforcement: II. The scalloped pattern of the cumulative record. *Journal of the Experimental Analysis of Behavior*, *29*, 67–75.

Dinges, C. W., Avalos, A., Abramson, C. I., Craig, D. P. A., Austin, Z. M., Varnon, C. A., … Wells, H. (2013). Aversive conditioning in honey bees (*Apis mellifera anatolica*): A comparison of drones and workers. *The Journal of Experimental Biology*, *216*, 4124–4134.

Doyen, S., Klein, O., Pichon, C.-L., Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown W., … Tucker, W.S. (1940). Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science*, *1*, 331–349.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York, NY: Appleton-Century-Crofts.

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)*, *17*, 69–77.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh, Scotland: Oliver & Boyd.

Gentry, G. D., Weiss, B., & Laties, V. G. (1983). The microanalysis of fixed-interval responding. *Journal of the Experimental Analysis of Behavior*, 39, 327–343.

Gigerenzer, G., (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gigerenzer, G., & Marewski, J. N., (2015). Surrogate Science: the idol of a universal method for scientific inference. *Journal of Management*, *41*, 421–440.

Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. San Diego, CA: Elsevier.

Grice, J. W., Craig, D. P. A., & Abramson, C. I. (2015). A simple and transparent alternative to Repeated Measures ANOVA. *SAGE Open*, *5*, 1–13.

Helmholtz, H. Von (1887). Numbering and measuring from an epistemological viewpoint. In P. Hertz, & M. Schlick (Eds.), *Hermann von Helmhotlz: Epistemological writings* (pp. 72–114). Dordrecht, Holland: Reidel.

Humphry, S. (2013). Understanding measurement in light of its origins. *Frontiers in Psychology*, *4*, 113.

Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, *115*, 288–299.

Izard, C., Dougherty, L., & Hembree, E. (1983). *A system for identifying affect expressions by holistic judgments (Affex)*. Newark, DE: Instructional Resources Center, University of Delaware.

Jaeger, K. M., Cox, A. H., Craig, D. P. A., & Grice, J. W. (2016). Auditory startle response predicts introversion: An individual analysis. *Modern Psychological Journal*, *23*, 67–78.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532,

Jordan, E. M., & Thomas, D. G. (2017). Contagious positive affective responses to laughter in infancy. *Archives of Psychology*, *1*, 1–21.

Laurent, E., & Lejeune, H. (1985). Temporal regulation of behavior in a fresh water turtle, Pseudemys scripta elegans (Wied). *Behavioural Processes*, *10*, 159–160.

Lejeune, H., & Richelle, M. (1982). Fixed-interval performance in turtle doves: A comparison with pigeons and rats. *Behaviour Analysis Letters*, *2*, 87–95.

Mace, C., & Kratochwill, T. (1986). The individual subject in behavioral analysis research. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 153–180). New York, NY: Plenum.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology* (3rd ed.). New York, NY: Chapman and Hall/CRC.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *Journal of Mathematical Psychology*, *38*, 244–273.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, England: Cambridge University Press.

Michell, J. (2003). The quantitative imperative, positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology*, *13*, 5–31.
Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*, 7–24.
Michell, J. (2014). Numbers as quantitative relations and the traditional theory of measurement. *The British Journal for the Philosophy of Science*, *45*, 389–406.
Michell, J., & Ernst C. (1997). The axioms of quantity and the theory of measurement: Translated from Part I of Otto Hölder's German Text "Die Axiome der Quantität und die Lehre vom Mass". *Journal of Mathematical Psychology*, *40*, 235–252.
Nosek, B., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.
Place, A. J., Varnon, C. V., Craig, D. P. A., & Abramson C. I. (2017). Exploratory investigations in operant thermoregulation in rattlesnakes (*Crotalus atrox*). In M. J. Dreslik, W. K. Hayes, S. J. Beaupre, & S. P. Mackessy (Eds.), *The biology of rattlesnakes II* (pp. 213–227). Rodeo, NM: ECO Herpetological Publishing and Distribution.
Schorske, C. E. (1997). The new rigorism in the human sciences 1940-1960. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation: Fifty years, four disciplines* (pp. 309–329). Princeton, New Jersey: Princeton University Press.
Schneider, B. A. (1969). A two-state analysis of fixed-interval responding in the pigeon. *Journal of the Experimental Analysis of Behavior*, *12*, 677–687.
Skinner, B. F. (1956). A case history in scientific method. *American Psychologist*, *11*, 221–233.
Skinner, B. F. (1972). *Cumulative record*. New York, NY: Appleton-Century- Crofts.
Solovey, M. (2004). Riding natural scientists' coattails onto the endless frontier: The SSRC and the quest for scientific legitimacy. *Journal of the History of the Behavioral Sciences*, *40*, 393–422.
Spiezio, C., Leonardi, C., & Regaiolli, B. (2017). Assessing colour preference in Aldabra giant tortoises (*Geochelone gigantea*). *Behavioural Processes*, *145*, 60–64.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.
Stevens, S. S., & Davis, H. (1938). *Hearing: Its psychology and physiology*. New York,NY: Wiley.
Thorngate, W., & Carroll, B. (1986). Ordinal Pattern Analysis: A strategy for assessing hypotheses about individuals. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 201–232). New York, NY: Plenum Press.
Thorngate, W., & Edmonds, B. (2013). Measuring simulation observation fit: An introduction to Ordinal Pattern Analysis. *Journal of Artificial Societies and Social Simulation*, *16*. Retrieved from http://jasss.soc.surrey.ac.uk/16/2/4.html.
Thorngate, W., & Ma, C. (2016). Wiggles and curves: The analysis of ordinal patterns. *Problemy Zarządzania*, *14*, 160–172.
Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*, 65–72.
Whissell, C., Abramson, C. I., & Barber, K. R. (2013). The search for cognitive terminology: An analysis of comparative psychology journal titles. *Behavioural Sciences*, *3*, 133–142.
Woolston, C. (2015). Psychology journal bans P values. *Nature*, *519*, 9.
Zeiler, M. D., & Powell, D. G. (1994). Temporal control in fixed-interval schedules. *Journal of the Experimental Analysis of Behavior, 61,* 1–9.

Financial conflict of interest: Support for this research was provided by grants NSF-REU (2016-1560389), and NSF-OISE (2015-1545803).
Conflict of interest: No stated conflicts.

Submitted: *January 13th, 2018*
Resubmitted: *March 24th, 2018*
Accepted: *April 7th, 2018*