MASTER THESIS

Learning in complex systems – Specialization in Behavior Analysis - MALKA

09 2015

Main title:

Stimulus Equivalence: Conceptual and Methodological Issues

Article 1:

Concepts: Worldviews, Cognitive Theories, and Stimulus Equivalence

Article 2:

The Sorting Test: On the Measurement of Equivalence Class Formation

Sjur Granmo

Faculty of Health Sciences

Department of Behavioral Science

OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES

Acknowledgement

There are four people and one family that deserve my gratitude and appreciation for all their help and understanding throughout these two years. First of all, I want to thank my supervisor Prof. Erik Arntzen who has been a true inspiration throughout my time in study, and he will continue to be one of my greatest sources of motivation in relevant work in the future. Thank you for your marvelous work and I can only hope for more collaboration in the future. Second, I want to thank Prof. Jon Arne Løkke and Gunn Elisabeth Haagensen Løkke who led me into the interest for psychology, science, and behavior analysis in the first place. I could not be more grateful for having you as good friends and colleagues. Your hospitality and willingness to help and support throughout my study have been magnificent. Third, I want to thank my family—my mother, father, and brother—who have given me an upbringing not comparable to any other: one that I will be outmost grateful for throughout my life. You have made this possible. At last, but until the end—Kjersti Bye—thank you.

Article 1

Concepts: Worldviews, Cognitive Theories, and Stimulus Equivalence

Sjur Granmo

Faculty of Health Sciences

Department of Behavioral Sciences

Abstract

The research field of concepts is—in the present article—divided in two accounts: the cognitive psychology approach and the behavior analysis approach. The article starts off by illuminating the philosophical assumptions related Pepper's (1942) world hypothesis, or worldviews, restricted to mechanism and contextualism. Later, there are drawn analogies between the worldviews and the scientific practices of cognitive psychology and behavior analysis. With respect to cognitive psychology, main characteristics of three of the theories of the structural study of concepts will be described. With respect to behavior analysis, the study of variables influencing concept formation, and stimulus equivalence, will be presented with links drawn to the early work of Murray Sidman on conditional discrimination training. The conclusion of the article will consist of some suggestions of systematic replications that are needed with respect to earlier research on the measurement of equivalence class formation. The article will end by pinpointing the general purpose of Article 2.

Keywords: concepts, cognitive psychology, behavior analysis, worldviews

After decades of scientific research, we have arrived at different definitions of *concepts* and *concept formation*. Although great differences within the discipline of cognitive psychology, there seem to be an agreement that concepts are subpropositional or structured mental representations (Margolis & Laurence, 1999). On the other hand, behavior analysts have defined concepts as generalization within and discrimination between stimulus classes (Keller & Schoenfeld, 1950). But it has been argued that concepts contain *meaning* and *references*, and they have symbolic value for other stimuli. The phenomena of *stimulus equivalence* brought with it the conceptualization of meaning and references. These different approaches will be illuminated in the present article. The present article will not try to give a review of the exhaustive literature regarding either of the approaches (i.e., cognitive psychology and behavior analysis). Rather, the present article will—at this issue—remain at the scope of describing the main characteristics within each discipline, and thereby some of the crucial differences between the disciplines.

The differences between the cognitive psychology and behavior analysis are mainly different approaches in the methodological and explanatory manners with regards to the data (e.g., Dougher, 1995). However, the approaches seem to be based on different assumption of the epistemology of the phenomena of interest. Such philosophical assumptions seem to be a reasonable approach when encountering differences in the understanding of any phenomena studied across scientific disciplines (e.g., Dougher, 1995; Leigland, 2003).

As an attempt to understand the philosophical core of the two disciplines that have encountered the scientific field of concepts, the purpose of the present article is to illuminate the basic assumptions of *mechanism* and *contextualism* proposed by Pepper (1942). More detailed, we will analyze some correspondences between these two worldviews and the conceptualization of the subject matter in general, and concepts in particular, in both cognitive psychology and

behavior analysis. We will start by addressing the worldviews, thereby being able to link the differences between scientific disciplines directly to mechanism and contextualism.

Note that the present article does not intend to compare the worldviews of mechanism and contextualism to the philosophy of any other kind that relates to cognitive psychology and behavior analysis, respectively (e.g., pragmatism). Moreover, by illuminating the characteristics of mechanism and contextualism and linking them to cognitive psychology and behavior analysis, we do not assume that the worldviews are synonymous or equal to other philosophies of either of the two disciplines (e.g., radical behaviorism).

The Worldviews

Pepper (1942) proposed four world hypotheses that would count for ways of looking at the nature of different phenomena in the world for scientific purposes. The four views are formism, mechanism, contextualism, and organicism. Although most of the world hypotheses, or worldviews, have some properties in common, they are distinguished for a reason. Each worldview consists of a *root metaphor* as well as a *truth criterion* along with other properties that defines their category membership. The present article will elaborate on such issues concerning the differences between mechanism and contextualism. We will look closer on the distinction between the two in the following paragraphs.

Mechanism

The root metaphor in mechanism is the *machine*. That is, when looking at a particular phenomenon and its relating events, mechanism makes use of the machine as a way of understanding how the phenomenon is organized. In the machine, the parts are related and organized a priori and fixed matter (Fox, 2008). There are often causal relationships between the

parts, for example, when a switch triggers the light bulb to illuminate, or when the ignition switch causes a chain of operations to occur before a motor start running. With respect to cognitive psychology, mechanism can be expressed as the conceptualization of discrete parts within the organism. The discrete parts can take many forms, but some common terms are the perception, storage, processing, retrieval or representation of information. These parts and their subsequent processes are often conceptualized as resulting in a response that is observable (Hayes, Hayes, & Reese, 1988).

The truth criterion of mechanism is *correspondence*. Hayes et al. (1988) writes, "…the knower knows a copy of the world, not the world itself. Truth is a matter of how well the copy corresponds to the world…" (p. 99). That is, correspondence between the world and the way we are describing the interrelated parts of it. The goal of the scientific practice is to discover the parts of the machine, as well as the relations among them.

Mechanism is the *analytic* and *integrative* by its properties. By analytic, the whole can be reduced to its parts. Mechanists views parts as interesting in themselves, and any relationship to other parts does not change their nature. That is, the parts are essential in nature, and the relations among parts do not define each part. Rather, the structure and properties of each part contains the necessary and sufficient attributes for the defining features. By integrative, we assume that facts are related before any systematic observation or experiences of the phenomenon of interest are conducted. A common practice is to test whether hypotheses can be verified or falsified on new phenomena (Fox, 2008; Hayes et al., 1988). If the verbal construction (i.e., prediction/hypothesis) corresponds highly to the real world phenomenon, then we have obtained the scientific goal of prediction verification and the truth by correspondence are met (Hayes et al., 1988; Morris, 1988). In final remarks, an important aspect of this implications in psychology is that one need

not to have seen, or by any means experienced, each part that are "placed in the organizational map". Unobservable parts, by principle, can be added to the formulae if this yields an appropriate understanding of the whole (e.g., the "retrieved representation" of a stimulus proposed in a recent paper by Craddock & Miller, 2014). However, we recommend Dougher (1995) for an example of a discussion on the differences in explanatory practices and the possible disagreements between mechanism and contextualism. Such discussions seem to be easily provoked when members of the different scientific disciplines are seemingly neglecting the philosophical core behind their scientific practices.

So, in summary of this brief presentation, mechanism uses the analogy of the machine and the truth by correspondence. Verbal constructions are made about the real world and new phenomena are implied by the construction (Hayes et al., 1988; Morris, 1988).

Contextualism

The root metaphor in contextualism is the "act in context", or the "historic event" (Pepper, 1942, p. 232). Both metaphors are illuminating how a contextualist considers the relationship among the events available for scientific investigation. A historical act may be misleading if perceived as something that has been and can be forgotten. The term historical is merely placing the emphasis on earlier interactions between the act and the context that may be necessary to include in a thoroughgoing analysis at the present time (Hayes et al., 1988).

Contextualism has the pragmatic truth criterion of *successful working*. The reader will notice that the term truth, although an important matter to be discussed, is sometimes slightly less emphasized when it comes to pragmatism (Barnes-Holmes, 2000). Truth in pragmatism seem contrasted by the search for a *real* truth in a world consisting of two dimension, one dimension that is the world as it is, and another dimension in which is our perception of it, by which often

are given ontologically rich elaborations. Pragmatists classify knowledge, or an idea, as true when such statements serve some kind of purpose or are of utility. The knowledge need not be the exact truth about the world, but it needs to be useful in achieving the goal of prediction and control (Fox, 2006; Leigland, 2003).

Seen from behavior-analytic grounds, the conceptualization of the 3-term contingency serves as a good example. When the scientific goal is to change behavior in order to understand it, we will have a rich experience of the phenomenon of behavior when the rate of responding in a food-deprived lever-pressing rat increases after repetitions of food delivery contingent upon such behavior. That is, by manipulating an independent variable (i.e., food) it yield changes in a measured dependent variable (i.e., the behavior of the rat); a functional relation. This is a fact only in so far as changing the environmental events alters the rat's behavior. There are no statements of absolute truth and reality in the demonstration of these functional relations, only a demonstration of an act in context where the contextual variables are controlling the act. It is also evaluated by the pragmatic truth criterion of successful working.

Contextualism is *synthetic* and *dispersive* (i.e., as opposed to the analytic and integrative mechanism). By synthetic, contextualism emphasizes that the whole is basic, whereas the parts are derived (Hayes et al., 1988). That is, phenomena are of interest by studying the whole act as one, not just each of the parts by itself. The ongoing act is never studied alone, but is seen in combination with the myriads of current and historical contextual events happening together with the act (Fox, 2006). For example, if reduced to an experimental chamber, we are concerned with the pressing of a lever as one part, but it would not serve the scientific purpose of control if we do not include the food-delivery that maintains the behavior. The basic whole can reduced to its parts only if it serves a practical purpose (e.g., distinguishing events by names). In other words,

contextualism emphasizes that the metaphor of the act-in-context necessarily has to include the contextual variables in the analysis as well as the act itself. By dispersive, contextualism does not assume that some facts or events are related, or stands in a causal relationship to one another, a priori of observation and experience. Instead, it requires empiric observation and experience and "facts are related when they are found to be so, not by assumption" (Hayes et al., 1988, p. 98).

In summary, contextualism is a dispersive study of an act in context in which meets the purpose of a synthesis, mechanism is an integrative study of the machine in which often leads to the structural analysis of the parts in relation to other parts.

Mechanism in Cognitive Psychology

As an trivial example, Murphy (2002) elaborates on how a concept might look like. He wrote that when we have *formed* a concept that corresponds to the category, at some later point, the concept would *help us understand* and, thereby acting appropriately upon, new members of the same category. Take a closer look at the statement. First of all, when we read that a concept has been formed, then many behavior analysts are questioning the procedural details of such formation. Second, the claim that the concept would be *formed by the person* and that it would *help the person at a later point* seem to deal with the inner states of a human organism; it has to be retrieved by the person of which had formed it. For behavior analysts, this is also a conceptualization where procedural details are missing. Moreover, since we cannot find out any more of the variables controlling the formation of concepts, we would be forced to end our analysis and such processes are argued to be impeding for the science of behavior (e.g., Skinner, 1953, 1968, 1974). Third—and here we see the sketches of mechanism—when talking about inner agents, here represented by mental representations, we easily include them in causal relationship by which the behavior of interest would never occur unless we conceptualized the

preceding mental activity. That is, mental representations (e.g., perception, storage, retrieval) are the part of the machinery chain by which elicit the necessary and observable response. However, this mechanistic conceptualization is not regarded as impeding for the science of cognitive psychology. Indeed, they are bound to invent this hypothetico construct with causal status as much as behavior analysts needs to invent the hypothetico construct of reinforcement. As Hayes et al., (1988) wrote,

> …contextualists might attack whatever force is invoked to explain the operation of the machine, arguing that the force is inferred from the events it "explains" and is, thus, an instance of reification. From a mechanistic standpoint, however, a driving force of some sort is necessary; it is not a matter of observation. (p. 106).

Hitherto, the fundamental differences between the two sciences have been described roughly. To get the greater impression of how it may look like when two opposing philosophies yields their own scientific practices and conceptualizations, we will now look at the research areas of concepts and concept formation. Concepts and the study of their structures and nature are presented with theories from cognitive psychology, whereas concept formation is presented with the conceptualization of stimulus classes and their emergent properties.

Concepts and Categorization in Cognitive Psychology

Classical Theory

In the classical theory, the concept is defined as having the necessary and sufficient conditions or properties. We may think of these properties as being listed up, and the properties give the opportunity to decide which to include and which to exclude from concept (e.g., legs, back, seat, physical, nonliving, something you sit in). The established properties in a concept

seem to be of a constant nature; to be regarded as essentialistic (Palmer, 2002). The process of inclusion and exclusion based on the properties or conditions are called encoding. Concepts are regarded as structured mental representations in which encode these properties so as to be applied in perceptual and sensory terms (Margolis & Laurence, 1999). Concepts are understood by collecting the concept's properties and a subsequent process of inclusion or exclusion of the sufficient properties in the concept. The process is called categorization and the properties to-be-categorized are called references of the concept.

Palmer (2002) reviewed the book of Margolis and Laurence (1999). From a behavior-analytic standpoint, he argues that the classical theory presupposes that one can identify examples of concepts. That is, we can determine if they are structured mental representations and whether they encode sets of sensory perceptual conditions. Of course, criticizing the cognitive field of research for presupposing or assuming facts before observation and experience, and doing so in order to fit a formulae or model, is like criticizing their philosophy of science, rather than their scientific practice. Much like Pepper (1942) warned us about interdisciplinary eclecticism, to criticize the assumptions based on another worldview and arguing that one worldview is more correct than the other is fruitless and confusing (see also Fox, 2008; Hayes et al., 1988). The presupposing property of the classical theory meets the integrative characteristics of a mechanistic model, and it is contrary to the dispersive contextualistic approach. However, both characteristics have their strengths and with respect to which of the worldview is the better, it depends on the purpose of the investigation, and thereby the scientific goal or objective (Fox, 2008).

Prototype Theory

The prototype theory holds that a statistical analysis of the properties that are most typical of a concept should yield important information about the concept's structure (Margolis & Laurence, 1999). So, the term concept is regarded as more of a distribution, rather than collection, of properties. The properties are regarded as more or less central or typical to the concept (e.g., robin may be more typical of a bird than is a swan; Palmer, 2002). This theory presents a less essentialistic assumption of concepts when it allows members to miss out on typical properties of a concept that indeed would be regarded as an instance close the features of a prototype. Further, the theory holds that we may see a tendency that two or more properties in some manner are correlating or at least occurs in some kind of consistency to one another. Contrary, as we saw in the classical theory, all properties needed to be present together (i.e., ultimate list).

There is a great deal of published papers by prototype theorists investigating the family resemblance (e.g., Rosch & Mervis, 1975) of the stimuli of a concept; rather than the more essentialistic and single-notion property attributed in the classical theory. One interesting aspect of the prototype theory is that the prototype itself need not be among the example; it can be an abstraction of the central properties that the subject has been exposed to (Palmer, 2002). In the prototype theory, the categorization process is regarded as a representation process in which we are representing the prototype before responding appropriately to the instance (Margolis & Laurence, 1999). That is, we compare the similarities of the abstracted and represented prototype (e.g., bird) and the instance or example we are considering (e.g., robin).

The view of the prototype theory seems more dynamic and less essentialistic than the classical theory, especially in that they include the fact that concept and their central properties changes with respect to experience, changes in society, and so on. Palmer (2002) writes: "One

might have expected the formulation of the prototype theory to lead to an inquiry into the variables that affect 'concept formation,' an inquiry that would necessarily touch, at least obliquely, on the basic behavioral processes." (p. 600). Even the prototype theory is grounded in mechanism and the characteristics of such a worldview, and as we have seen, one cannot expect the cognitive psychologists to "play by the rules" of the behavioral science. However, Palmer (2002) points out the exceptional work of Rosch and Mervis (1975) and Rosch (1978), and their research on prototypicality in relation to "…response rate, rate of acquisition, priming effects, and other measures of response strengths." (p. 600). These are useful investigations for a behavior analyst, of course, with respect to the pragmatic purpose of control or influence on the variables that operates on the phenomenon of concept formation.

Theory-Theory

Whereas the classical theory and the prototype theory emphasized that lists of suggested properties were collected or distributed along with statistical analyses, respectively, the theory-theory assumes that theoretical terms within psychology can be subject for philosophical treatment. This theory assumes that people ascribe and classify instances into concepts in an essentialistic manner. It also assumes that the conceptual changes of humans can be explained in line of theoretical changes in science. That is, thinking similar to scientific methodology is a crucial analogy to the way we conceptualize, interact, and learn to understand the world around us (Margolis & Laurence, 1999).

Similar to the classical theory and prototype theory, the theory-theory considers the role of a concept to be regarded as a description. The description contains many instances by which is considered to be referents of the concept. The theory assumes that people uses a mental representation theory that distinguishes and classifies the inner hidden properties from outer

observable properties in different concepts. For example, Gelman and Wellman (1991) investigated the internal and external changes to, for example, a dog. The children were exposed both to the imagination that a dog would have removed its insides, and that the dog would have removed its fur. The most of the children regarded the latter as still being a dog (i.e., emphasizing the internal properties as the constant/essential measure of a dog), whereas the former were not considered a dog. As this example shows, the theory-theory suggests that people have a tendency to categorize things as if the things had essential properties.

The essential properties of concepts are the main in objection made by Palmer (2002) in his review of the book of Margolis and Laurence (1999). In addition, there are many objections to a pure theory-based understanding of the phenomenon of concepts. Another example, also related to essentialism, may be Stanowich (2010) who wrote: "The refinement of conceptual terms comes from the interplay of data and theory that is inherent in the scientific process, not from debates on language usage" (p. 38).

In summary, the theories of concepts and categorization bring with them the characteristics of mechanism in general. The concepts are generally explained by the verbal statement of the structured mental representations, from which its parts (e.g., features) are consistent with the outcomes of different categorization tasks.

Contextualism in Behavior Analysis

Contextualism in behavior analysis is—in the present article—referring to *functional* contextualism (e.g., Fox, 2008), not *descriptive* contextualism. Mechanistic explanations such as the retrieval of mental representation as a cause for the behavior of interest are generally dismissed in behavior-analytic literature. In cognitive psychology, however, we see such utterances frequently; and why should we not? As part of the worldview of mechanism—and the

truth criterion of correspondence between a verbal statement and the perceived reality—cognitive psychologists are doing well and good. Behavior analysts need not accuse them on methodological and explanatory grounds. Now, we will give the readers a brief presentation of how contextualism is complementary with the well-evolved conceptual field of behavior analysis.

In the case of the mechanistic approach with the mental representation, the variable with causal status is considered to be the preceding event (i.e., cognitive process) before the behavior are emitted, or perhaps even elicited. Behavior analysts would argue against the practice of "placing" (i.e., as in a model or theory) a hypothetical causal mechanism in between two readily observable events. We would be determined to look for detectable variables that are operating in the environment (Baum, 2005).

In behavior analysis, one look for consistency between the environmental events and the behavior of an organism, and we needed no longer search for the internal variables of which have been assumed to cause the behavior. Skinner's contributions led to a thoroughgoing new philosophy of science called *radical behaviorism* (e.g., Skinner, 1938, 1953; 1957, 1974, 1981). We shall see that this philosophy have much in common with the contextrual worldview presented by Pepper (1942).

First of all, behavior analysts have accused cognitive psychologists for being essentialistic (Donahoe, 2003; Donahoe & Palmer, 1994; Palmer, 2002; Palmer & Donahoe, 1992). Essentialism is "…the idea that the only good scientific theories are those that give ultimate explanations of phenomena in terms of their underlying essences or their essential properties" (Stanowich, 2010, p. 37). That is, they tend to construe new variables that are reified in order to fit the model of explanation. However, incompatible to essentialism, Donahoe and Palmer have elaborated on the explanatory model of selectionism in behavior analysis (first introduced

through Skinner, 1981). In essentialism, variation is regarded as noise or interferences in an otherwise orderly and categorically world (cf. the integrative feature of mechanism). In selectionism, however, variation is considered to be a necessary part of the material on which selection is operating; thereby considering variation to be inevitable (Donahoe, 2003).

In radical behaviorism, the two variables that are of interest are the environmental events and the behavioral events. Behavior analysts treat these variables in a dynamic manner. Behavioral variation is accepted, and we measure the behavior in relation to the environment so as to see the consistency between the events. However, the behavior on one hand, and the environment on the other, are not considered important in their own right; it is the relation between what happens in the environment and what happens in the behavior that are of interest (the whole is basic and the parts derived in Hayes et al., 1988). More specifically, the stimulus change in the environment that happens *after* some variant of the behavioral variation have occurred is the variable that seem crucial: the consequence. The consequence is the environmental variable that are said to select the behavioral variant and when this operation results in higher frequency of the behavioral variant (i.e., the response), then we say that the behavior had been reinforced, the process are called reinforcement and such behavior is called operant behavior. For example, behavioral variation is much valued in the process of shaping. Imagine the food-deprived rat is sniffing around in the chamber and no food is presented yet. If we had an operational definition on sniffs, grinds, stretches, and so on, we would have about an equal amount of responses registered in a given interval of time. This behavioral activity was called the flat generalization gradient (Catania, 2007). The activity is the variation on which the selection by consequences (i.e., food-delivery) will operate. Now, imagine we start to present food at the stretching toward the ceiling. This behavior will be reinforced and thereby increase in

frequency. This operation is forcing the other behavioral events to decrease. However, we will still see a fair amount of the other responses occurring as well, but there is generally longer time between the sniffing responses relative to the stretching responses. The variation is reduced when we—in the procedure—was reinforcing one particular behavior of the rat. More importantly, although still variation left, it is not considered to be interfering, but to be necessary to change behavior even further.

So, a response occurs because it has in the history of the individual been followed by consequences, and when these consequences have followed the behavior, we can see that the behavior increased in frequency (i.e., the process of reinforcement), while other behaviors have not been increased. Although considered extremely complex and always a case of multiple causation, one trust that the consequential environmental variables are the causes of the behavior in a particular instance. Simply put, the behavior would not have occurred if the consequence were removed. If we removed the consequence, we would see the diminishing tendency in the pattern of responding. If it was reintroduced, we would se the increase again. Such a process can be repeated throughout many ABABABAB phases, and the controlling variable(s) are demonstrated over and over again. It also yields a good example of the successful working truth criterion, which pragmatism and contextualism shares, and it leaves us—as scientists—to respond effectively on such phenomena. By influencing behavior, we may understand its nature, although never claiming to know any absolute truth about it (Leigland, 2003).

So when cognitive psychologists dismiss behavior analysis as an account that can explain various psychological phenomena (Margolis & Laurence, 1999), one may ask how psychology is defined. The Oxford dictionary (OED, 2015) defines it as “The scientific study of the human mind and its functions, especially those affecting behavior in a given context”. Such a definition

seems biased with the worldview of mechanism and does not allow behavior analytic practice to be included into the field of psychology. On the other hand, imagine the behavior analyst accusing the cognitivist for not being able to identify the necessary observable and manipulable variables and thereby demonstrating the controlled changes in the dependent variable as a result of altering the independent variables (e.g., Skinner, 1953). By now, we hope that all the examples of arguing hitherto are starting to get satiating. We should be able to look through the cognitive vs. behavioral quarrel at this point and accept that the two scientific disciplines have very little in common other than their phenomena of interest. Philosophical assumptions such as the worldviews proposed by Pepper (1942) seem to be more of a personal choice than an actual fact (Fox, 2008).

On the other hand, we may be able to make use of the distinction between the different assumptions. For a behavior analyst, cognitive psychologists seem to describe the surface of the phenomenon of concept and neglecting the learning history that caused the concepts to enter the person's repertoire in the first place. That is, they seem to study how people are practicing concepts based on what they have learned in their lives up to that particular point when they received a particular task. While the goal of cognitive psychologists in concept research is to understand the mental representations that is assumed to exist, the behavior analysts—in accordance with their truth criterion—are trying to identify variables that are influencing, as well as establishing, the phenomenon (Arntzen, 2004; Arntzen, Grondahl, & Eilifsen, 2010; Arntzen & Holth, 1997). For example, Arntzen (2012, Table 2) lists a wide range of variables to be manipulated in the formation of equivalence classes.

Murray Sidman and Conditional Discrimination

Hitherto, we have devoted much elaboration on one crucial difference between cognitive psychology and behavior analysis, namely, the role of selection by consequences. It has been made clear that behavioral events do not happen in a vacuum, but is in fact followed by stimulus changes after it occurs. However, we have not looked into the emphasis that behavior analysts put on the antecedent event, other than the case of food-deprivation. Now we shall look at stimulus changes that happen before a response is emitted, and we will focus on discrimination training, and conditional discrimination training.

Early Work

Ever since Skinner published his work on response patterns during extinction trials (Skinner, 1933a), he pawed the way for over a half century of research on discrimination training. He was concerned with the reinforcement of responses in the presence of one stimulus, and a concurrent extinction of responses in the absence of that stimulus. In addition to develop patterns of discrimination, he also found that an already established discrimination could be revered by changing the contingencies with respect to the stimuli (e.g., S^D = light-off, S^Δ = light-on; Skinner, 1933b). Stimulus control, of course, being the changes in operant behavior shown in correlation with a stimulus that yield higher probability for responding (S^D), and one stimulus yield lower probability (S^Δ). The former stimulus' function has been altered after responding in its presence has produced the reinforcer, whereas responding in the presence of the latter stimulus has not produced the reinforcer.

In the 1960's, Murray Sidman worked on a hospital in Massachusetts and trained particularly people with different types of brain damages (e.g., aphasia) and people who were developmentally disabled. Most of the training was what has been called conditional discrimination training where the stimuli to-be-discriminated was circles and ellipses (Sidman &

Stoddard, 1966). In a series of article published throughout the 60's and 70's, data was collected on the effectiveness of fading procedures, in particular, to establish well-adjusted discriminations between small, but crucial changes in stimuli (e.g., Leichester, Sidman, Stoddard, & Mohr, 1969; Rosenberger, Mohr, Stoddard, & Sidman, 1968; Sidman & Rosenberger, 1967; Sidman & Stoddard, 1967; Stoddard & Sidman, 1967).

Mainly, the conditional discrimination procedure used by Sidman and colleagues consists of the 3-term contingency, but with an extra term added (Sidman, 1987). The 4-term contingency consists of a *conditional stimulus* (S^C), a discriminative stimulus (S^D), a response (R), and a reinforcer (S^R). The procedure would be best illustrated with the typical Identity Matching-To-Sample (ID-MTS) format. Lets assume that we have three stimuli that need to be discriminated between; blue, red, and green. The conditional stimulus may also be called the sample, and when the sample stimulus is presented, the participant will emit an observing response toward the sample stimulus. That is, blue is presented (S^C) and the observing response is emitted toward blue. When the observing response is emitted, the comparison stimuli will appear. The comparison stimuli consists of blue, red, and green; S^D, S^Δ, S^Δ, respectively, for that particular trial. For every conditional stimulus we have in the to-be-trained stimulus-stimulus relations we must have equally many S^D to-be-compared. So when red is the sample stimulus, then red is S^D, and the blue and green is the S^Δs, and so on.

Circle-ellipse discrimination.

After all necessary pre-training was completed, Sidman and Stoddard (1966) presented trials with the presentation of nine square matrices where the sample stimulus (i.e., a circle) was presented in the center of the matrices. After the observing response was emitted to the sample stimulus, then two stimuli were presented in two of the matrices around the center. One of the

stimuli displayed an exact copy of the sample stimulus in the center matrix and responding toward it would produce the programmed reinforcers. The other was a sharp-angled ellipse and responding toward it would only produce an intertrial interval. With such training, the participant would start to emit the correct response toward the circle in the outer matrices given the presence of the circle in the center matrix. After mastery performance in accordance with a given criterion, Sidman and colleagues started either (1) to fade in the appearance of other stimuli in the remaining matrices (i.e., ellipses of various angles), or (2) to fade the already discriminated relation between the full circle and the sharp-angled ellipse to become more similar to each other (i.e., making the ellipse more round).

Contextualistic features.

Now, we can see that—with inclusion of motivating operations (e.g., deprivation)—all necessary contextual variables are included in the explanation of performances that indeed has been defined as concepts (see Keller & Schoenfeld, 1950). We can also see that all conceptualization of stimulus class formation is in accordance with the pragmatic criterion of truth (i.e., understood by influence) and that the act-in-context metaphor are being supported with every contextual variable included. Now, we shall see a case in which conditional discrimination led to the phenomenon of stimulus equivalence; an example of stimulus generalization not explained by simple discrimination and generalization alone.

Sidman (1971)

In an attempt to establish what Sidman called reading comprehension he pawed the way for a field of research known as stimulus equivalence. This field would have as its primary objective to explain the variables controlling performances that was in accordance with the contingencies between stimulus-stimulus relations that was based on a limited set of trained

relations, where as the tested—and emerged relations—never had been directly trained. Reading comprehension was demonstrated when a person could correctly match the vocal pronunciation of the words (i.e., auditory stimulus) or the written word (i.e., visual stimulus) to the picture of the corresponding object or event (Sidman, 1971). That is, if the participant could the written word to the picture of the corresponding stimulus in the same class, he would have demonstrated reading comprehension. Thereby, if presenting the participant with the written word cat is in the presence of the picture of the cat, the participant would be able to point to the picture and thereby "understanding" that the picture of a cat and the written word *cat* is equal to each other in that particular context. Note that the relation between such different response modalities are arbitrary, in that they do not possess features that make them more probable to be grouped together (e.g., physical similarity). The only way a relationship between the stimuli may be established is by experiencing which of the stimuli that "belongs together" after receiving reinforcement after correct responding across sample and comparison stimuli.

The participant in the study was an "…institutionalized 17-year-old boy, microcephalic, and severely retarded." (p. 7). Before starting the conditional discrimination training, the subject could already match spoken auditory words to visual pictures of the corresponding object or event. We can call this relation for the AB-relation. In addition, the participant could match the visual picture to the spoken out word of the object or event by naming the stimuli he would see on the picture (i.e., a tact). The question of the experiment was if the training of mastery performance in matching spoken word to the visual written word (AC-relation) was enough to demonstrate the emergence of untrained relations between the trained stimuli. That is, Sidman (1971) tested if the boy could match the correct visual picture to the corresponding visual written word, when presented with the picture as sample stimulus (i.e., the untrained BC-relation). He

also tested if the boy could do the opposite; match the visual written word to the corresponding visual picture (i.e., the untrained CB-relation). At last he demonstrated that the newly learned relation to the visual written word could be tacted in the same manner as with the picture and the spoken-to-him word.

Summarized, the boy first learned to match auditory words to visual pictures, and then to visual written words later. These two relations between the stimuli seemed to be sufficient prerequisites for the emergence of both the untrained picture-word relation, and the word-picture relation. Note that this study used One-To-Many (OTM] training structure where the A-set of stimuli as sample was trained to the B-set of stimuli as comparisons, before the A-set again is the sample stimuli that was trained to the C-set of stimuli. The Linear Series (LS) training structure is designed to demonstrate all the emergent relations by training A to B, then B to C. Now, we have the opportunity to test for symmetry relations in BA and CB, and transitive relations in AC, as well as the equivalence relation in CA. However, as seen in the OTM structure, the BC and CB relations are the relation that is called transitive and equivalence relations (e.g., Arntzen & Holth, 1997).

Concepts as 4-Term Contingency Operants

To make sure the reader follows the procedure of conditional discrimination, we might as well describe the procedure of the training of baseline relations in a typical MTS format stimulus equivalence experiment. Let us assume that we are forming three classes of stimuli in a LS training structure. Each stimulus within each class are named by the letters A, B, and C. Therefore, in stimulus set A, there is A1, belonging to Class 1; A2, belonging to Class 2; and A3, belonging to Class 3. The same would be true for stimulus set B and C. That is, Class 1 consists of A1, B1, and C1; Class 2 consists of A2, B2, and C2; whereas Class 3 consists of A3, B3, and

C3. By now, we see that we have enough classes and enough members to start the conditional discrimination training (i.e., minimum two classes, but see Sidman, 1987). Since we use a LS training structure, we will be training all the A-stimuli to all the B-stimuli. This means that all A-stimuli will be sample stimuli when the B-stimuli are comparison, and no other order will be trained. We will also train all the B-stimuli to all the C-stimuli, meaning that all B-stimuli will be sample stimuli when all C-stimuli are comparisons.

For the convenience, let us assume that a participant did not discriminate between the mathematical sizes such as percentage, fraction, and the picture of a corresponding part of a circle (cf. Lynch & Cuvo, 1995) prior to the experiment. In the case of letters (i.e., members) and numbers (i.e., classes), Class 1 can be called “quarter”, Class 2 can be called “half”, and Class 3 can be called “three-quarters”. Note that the names just given to the classes are not among the to-be-trained baseline relations. In each stimulus set, we will give the members properties that are arbitrary to the other members of the same class. That is, A1, A2, and A3, will be given the property of percentage (i.e., 25 %, 50 %, and 75 % respectively). In the second stimulus set, B1, B2, and B3, would be 1/4, 2/4, and 3/4, respectively. In the third stimulus set, C1, C2, and C3 would be the pictures of a quarter of a circle, a half of circle, and three quarters of a circle. Now the participant starts the training, and he or she will receive feedback on correct responses whenever they press the corresponding comparison stimulus on each occasion as a sample stimulus of the same class are presented. The trained relations are AB and BC in each class. That is, although the order is randomized from trial to trial, the following trials would be repeated until mastery levels: A1 to <u>B1</u>, B2, B3 (correct response is underlined); A2 to B1, <u>B2</u>, B3; A3 to B1, B2, <u>B3</u>; B1 to <u>C1</u>, C2, C3; B2 to B1, <u>B2</u>, B3; B3 to C1, C2, <u>C3</u>.

After the participant scores by the designated mastery levels (e.g., 100 %) in training, the test for emergent relations can be administered. Here we are testing for the properties that define stimulus equivalence in extinction trials (e.g., Sidman, 1992). Remember that concepts are defined only as the stimulus classes in which we can document generalization within classes of stimuli, and the discriminations between one class from another (Keller & Schoenfeld, 1950). However, stimulus equivalence classes—although still stimulus classes—would be able to document the properties of reflexivity, symmetry, and transitivity. Reflexivity test trials would be mastered if the participant responded in A1 of all A-stimuli when presented with A1 as a sample. The same would be necessary to document the remaining reflexivity relations (e.g., A2 with A2, etc.). Symmetry would be demonstrated by presenting, for example, B1 as a sample and present A1, A2, and A3 as comparisons, and of course, in the same manner with all remaining symmetry relations (i.e., BA and CB relations). Transitivity would be documented when, for example, A1 was presented as a sample stimulus and the C-stimuli (i.e., C1, C2, and C3) were presented as comparisons, and in the same manner with the remaining transitivity relations (A2 and A3 to the C-stimuli). An additional tested relation is included in the combined symmetry-transitivity test for the emergent relation called the equivalence relation (e.g., Sidman, 1992).

Now, it does not matter whether we present the participant with fractions as sample and percentage as comparison; or the picture of half a circle as sample and percentages or fractions as comparisons for that matter. The stimuli relations that have been trained are now interchangeable, and the performance have been established only by training the few conditional baseline relations of AB and BC relations. Interchangeable because we may have the just as high probability for a correct response independent of which stimuli we present as samples and comparisons. We may say that the participant “understands” some of the features of the concept “fractions”,

"percentage", and "parts of a circle"—as well as the more general concept of "sizes". In addition, we know that the members of each class are interchangeable to one another in specific contexts, in the same way as Sidman (1971) found the emergent relations in the boy learning to read with comprehension. Sidman and Tailby (1982a) elaborates best in their own words on this issue from the 1982 article on the stimulus equivalence paradigm:

> Pointing to a picture in response to a printed word denotes reading comprehension only if the word and picture are related by equivalence and not merely by conditionality. Stimulus classes formed by a network of equivalence relations establish a basis for referential meaning. The equivalence paradigm provides exactly the test that is needed to determine whether or not a particular conditional discrimination involves semantic relations. Linguistic analysis has challenged functional behavioral analysis to account for new behavior that has no apparent reinforcement history (e.g., Chomsky, 1965; Fodor, Bever, & Garrett, 1974). The equivalence paradigm takes a short step in this direction by specifying procedures for generating new and seemingly unreinforced matching to sample and oral naming. (p. 20).

After establishing equivalence relations, we are left with what Sidman (1986) termed as four-term contingencies. In short, such contingencies—after necessary training—allow that one stimulus will occasion responding to a discriminative stimulus. In other words, depending on the sample stimulus the participant is presented with, the participant will respond to the discriminative comparison stimulus it occasions. For example, the discriminative comparison stimulus, B1, which is occasioned by the conditional sample stimulus A1 becomes a delta stimulus if another stimulus, say A3, was presented as sample stimulus in the next trial. This will be the case also for the remaining combination of A2 as sample stimulus and B1, B2, and B3 as

comparison stimuli. Also, five-term contingencies were proposed by Sidman (1986, p. 239) to describe "…the influence which the environment exerts over conditional discriminations". For example, a given tone could occasion the selection of a specific hue, which was conditional upon the selection of a specific form. Note that Sidman's emphasis on the influence of the environment in his statement above is directly related to the root metaphor, as well as the truth criterion, in contextualism.

Future Directions

Characteristics of Equivalence Classes

Stimulus equivalence research has been evolving in the last few decades. Besides the already listed manipulable variables and parameters, suggested by and documented in Arntzen (2012), the following characteristics have been shown to be characteristic of equivalence relations.

First, when retesting the emergent relations within each of the classes after 2–5 months in extinction trials, participants have demonstrating that the emergent relations were maintained still (Saunders, Wachter, & Spradlin, 1988). Second, equivalence classes are readily receptive to new members. That is, new members can be added to the class in both the same modality and with different modalities (Saunders et al., 1988). Third, Saunders, Saunders, Kirby, and Spradlin (1988) showed that equivalence classes are resistant to change, and that a history of equivalence class emergence alone was "sufficient to produce class merger and class development when the subject is exposed to conditional discrimination problems of a similar type" (p. 160).

Fourth, when some members of an already established equivalence class have been trained and subsequently been altered in function (e.g., discriminative for clapping, waving, fast

or slow responding), then this effect has been generalized to the remaining members of their respective, and are now well-known properties of equivalence classes (Transfer og Function; see for example Barnes & Keenan, 1993; Hayes, Devany, Kohlenberg, Brownstein, & Shelby, 1987; Lowe, Horne, & Hughes, 2005).

Fifth, the transfer of function has also been demonstrated with respect to classical conditioned behaviors (i.e., respondents). Dougher, Augustson, Markham, Greenway, and Wulfert (1994) established two 4-member equivalence classes (i.e., A1, B1, C1, and D1 in Class 1; A2, B2, C2, and D2 in Class 2). Then, they altered the effect of B1 by presenting the B1-stimulus with a small electric shock. The respondents were recorded by galvanic skin response measurement, and when the participants were presented with any of the stimuli from Class 1, galvanic skin responses were elicited, but they were not elicited by the presentation any of the members of Class 2 (i.e., where no members had been classically conditioned).

The research hitherto has explored important characteristics relating to the variables responsible for the formation of equivalence classes. Of course, there are different explanations (Hayes, Barnes-Holmes, & Roche, 2001; Horne & Lowe, 1996; Sidman, 1990, 1992, 1994, 2000), but relatively small changes are done with respect to subtle parameters from experiment to experiment, and such systematic replications are considered very effective in demonstrating the generality of the phenomenon (Sidman, 1960).

Systematic Replications Needed

The generality of the phenomenon of stimulus equivalence can easily be questioned. More than 1000 articles have studied equivalence classes by investigating the manipulable variables that are directly responsible for the formation of such classes (Arntzen, Norbom, & Fields, 2015). However, the formats in which the training and testing of the class formation have demonstrated

have been predominantly been shown by the MTS format. Only a few exceptions have been seen. For example, Fields, Doran, and Marroquin (2009) showed that the phenomenon of equivalence could be generalized over the training and testing phases of stimulus pairing (i.e., Y or N trials). Eikeseth, Rosales-Ruiz, Duarte, and Baer (1997) have shown that the baseline relations can easily be established by instructions or rules and they also used a pencil-and-paper format. This was replicated and extended by Smeets, Dymond, and Barnes-Holmes (2000). Research are needed to explore the generality of the phenomenon, and this can be achieved partly by using different training and testing formats (Arntzen et al., 2015; Dymond & Rehfeldt, 2001; Fields, Arntzen, & Moksness, 2014; Fields et al., 2009; Fields, Reeve, Varelas, Rosen, & Belanich, 1997). Many of the mentioned studies have used formats that are more or less based on forms of categorization. Among with response latency, or reaction time, verbal reports, and stimulus recall, Dymond and Rehfeldt (2001) suggested stimulus sorting as a supplementary measure of stimulus class formation. To include stimulus sorting as a test for class formation, it needs to be compared and evaluated along with the predominantly used MTS test format. The purpose of Article 2 is to critically evaluate the sorting test for class formation and thereby contribute to the rapidly evolving literature on different training and testing formats to be compared with the MTS test format.

References

Arntzen, E. (2004). Probability of equivalence formation: Familiar stimuli and training structure. *The Psychological Record, 54*, 275–291.

Arntzen, E. (2012). Training and testing parameters in formation of stimulus equivalence: Methodological issues. *European Journal of Behavior Analysis, 13*, 123–135. Retrieved from http://www.ejoba.org.

Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The effects of different training structures in the establishment of condiitonal discriminations and subsequent performance on tests for stimulus equivalence. *The Psychological Record, 60*, 437–462.

Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record, 47*, 309–320.

Arntzen, E., Norbom, A., & Fields, L. (2015). Sorting: An alterternative measure of class formation. *The Psychological Record, 65*(2), 1–11. doi: 10.1007/s40732-015-0132-5

Barnes, D., & Keenan, M. (1993). A transfer of function through derived arbitary and nonarbitrary stimulus relations. *Journal of the Experimental Analysis of Behavior, 59*, 61–81. doi: 10.1901/jeab.1993.59-61

Barnes-Holmes, D. (2000). Behavioral Pragmatism: No place for reality and truth. *The Behavior Analyst, 23*, 191–202.

Baum, W. M. (2005). *Understanding Behaviorism*. Malden, MA: Blackwell Publishing.

Craddock, P., & Miller, R. R. (2014). Attention as an acquisition and performance variable (AAPV). *Learning & Behavior, 42*, 105–122.

Donahoe, J. W. (2003). Selectionism. In K. A. Lattal & P. N. Chase (Eds.), *Behavior theory and philosophy* (pp. 103-128). New York: Kluwer Academic / Plenum Publishers.

Donahoe, J. W., & Palmer, D. C. (1994). *Learning and Complex Behavior*. Needham Heights, MA: Allyn and Bacon.

Dougher, M. J. (1995). A bigger picture: cause and cognition in relation to differing scientific frameworks. *Journal of Behavioral Therapy & Experimental Psychiatry, 26*(3), 215–219.

Dougher, M. J., Augustson, E., Markham, M. R., Greenway, D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior, 62*, 331–351. doi: 10.1901/jeab.1994.62-331

Dymond, S., & Rehfeldt, R. A. (2001). Supplemental measures of derived stimulus relations. *Experimental Analysis of Human Behavior Bulletin, 19*, 8–12.

Eikeseth, S., Rosales-Ruiz, J., Duarte, A., & Baer, D. M. (1997). The quick development of equivalence classes in a paper-and-pencil for- mat through written instructions. *The Psychological Record, 47*, 275–284. Retrieved from: http://opensiuc.lib.siu.edu/tpr/vol247/iss272/277.

Fields, L., Arntzen, E., & Moksness, M. (2014). Stimulus Sorting: A quick and sensitive index of equivalence class formation. *The Psychological Record, 64*, 487–498.

Fields, L., Doran, E., & Marroquin, M. (2009). Equivalence class formation in a trace stimulus pairing two-response format: Effects of response labels and prior programmed transitivity induction. *Journal of the Experimental Analysis of Behavior, 92*, 57–84.

Fields, L., Reeve, K. F., Varelas, A., Rosen, D., & Belanich, J. (1997). Equivalence class formation using stimulus-pairing and yes-no responding. *The Psychological Record, 47*, 661–686. Retrieved from: http://search.proquest.com/openview/646adf669bbd595436b501475a595216a595428dbb595430/595431?pq-origsite=gscholar.

Fox, E. J. (2006). Constructing a pragmatic science of learning and instruction with functional contextualism. *Educational Technology Research and Development, 54*, 5–36. doi: 10.1007/s11423-006-6491-5

Fox, E. J. (2008). Contextualistic perspectives. In J. M. Spector, M. D. Merrill, J. v. Merriënboer & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3 ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition, 38*, 213–244.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame Theory: A post-Skinnerian account of human language and cognition*. New York, NY: Kluwer Academic/Plenum.

Hayes, S. C., Devany, J. M., Kohlenberg, B. S., Brownstein, A. J., & Shelby, J. (1987). Stimulus equivalence and the symbolic control of behavior. *Revista Mexicana de Análisis de la Conducta, 13*, 361–374.

Hayes, S. C., Hayes, L. J., & Reese, H. W. (1988). Finding the philosophical core: a review of stephen c. Pepper's world hypothesis: a study in evidence. *Journal of the Experimental Analysis of Behavior, 50*(1), 97–111.

Horne, P. J., & Lowe, C. F. (1996). On the origins of naming and other symbolic behavior. *Journal of the Experimental Analysis of Behavior, 65*(1), 185–241.

Keller, F. S., & Schoenfeld, W. N. (1950). *Principles of psychology: A systematic text in the science of behavior*. New York, NY: Appleton-Century-Crofts.

Leichester, J., Sidman, M., Stoddard, L. T., & Mohr, J. P. (1969). Some determinants of visual neglect. *Journal of Neurology, Neurosurgery & Psychiatry, 32*, 580-587.

Leigland, S. (2003). In Response; Is a new version of philosophical pragmatism necessary? A reply to Barnes-Holmes. *The Behavior Analyst, 26*, 297–304.

Lowe, C. F., Horne, P. J., & Hughes, C. J. (2005). Naming And Categorization In Young Children: III. Vocal Tact Training And Transfer Of Function. *Journal of the Experimental Analysis of Behavior, 83*, 47–65. doi: 10.1901/jeab.2005.31-04

Lynch, D. C., & Cuvo, A. J. (1995). Stimulus equivalence instruction of fraction-decimal relations. *Journal of Applied Behavior Analysis, 28*(2), 115–126.

Margolis, E., & Laurence, S. (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.

Morris. (1988). Contextualism: The world view of behavior analysis. *Journal of Experimental Child Psychology, 46*, 289–323.

Murphy, G. L. (2002). *The Big Books of Concepts*. Cambridge, MA: MIT.

OED. (2015). Oxford Engish Dictionary; Definition of Psychology in English. http://www.oxforddictionaries.com/definition/american_english/psychology

Palmer, D. C. (2002). Psychological essentialism: A review of E. Margolis and S. Laurence (eds.), Concepts: Core Readings. *Journal of the Experimental Analysis of Behavior, 78*, 597–607.

Palmer, D. C., & Donahoe, J. W. (1992). Essentialism and selectionism in cognitive science and behavior analysis. *American Psychologist, 47*, 1344–1358. doi: 10.1037/0003-066X.47.11.1344

Pepper, S. (1942). *World hypothesis: A study in evidence*. Berkely, LA: University of California Press.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Rosenberger, P. B., Mohr, J. P., Stoddard, L. T., & Sidman, M. (1968). Inter- and intramodality matching deficits in a dysphasic youth. *The Journal of the American Medical Association, 18*, 549–562.

Saunders, R. R., Saunders, K. J., Kirby, K. C., & Spradlin, J. E. (1988). The merger and development of equivalence classes by unreinforced conditional selection of comparison stimuli. *Journal of the Experimental Analysis of Behavior, 50*, 145–162. doi: 10.1901/jeab.1988.50-145

Saunders, R. R., Wachter, J., & Spradlin, J. E. (1988). Establishing auditory stimulus control over an eight-member equivalence class via conditional discrimination procedures. *Journal of the Experimental Analysis of Behavior, 49*, 95–115. doi: 10.1901/jeab.1988.49-95

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books, Inc.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of Speech and Hearing Research, 14*, 5–13.

Sidman, M. (1986). Functional analysis of emergent verbal classes. In T. Thompson & M. D. Zeiler (Eds.), *Analysis and integration of behavioral units* (pp. 213–245). Hillsdale, NJ: Lawrence Erlbaum.

Sidman, M. (1987). Two choices are not enough. *Behavior Analysis, 22*, 11–18. Retrieved from http://www.equivalence.net/pdf/Sidman_1987.pdf.

Sidman, M. (1990). Equivalence relations: Where do they come from? In D. E. Blackman & H. Lejeune (Eds.), *Behaviour analysis in theory and practice: Contri- butions and controversies* (pp. 93–114). Hillsdale, NJ: Erlbaum.

Sidman, M. (1992). Equivalence relations: Some basic considerations. In S. C. Hayes & L. J. Hayes (Eds.), *Understanding verbal relations (pp. 15–27)*. Reno, NV: Context Press.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior, 74*(1), 127–146.

Sidman, M., & Rosenberger, P. B. (1967). Several methods for teaching serial position sequences to monkeys. *Journal of the Experimental Analysis of Behavior, 10*(5), 467–478.

Sidman, M., & Stoddard, L. T. (1966). Programming perception and learning for retarded children. In N. R. Ellis (Ed.), *International review of research in mental retardation* (Vol. 2). New York, NY: Academic Press.

Sidman, M., & Stoddard, L. T. (1967). The effectiveness of fading in programming a simultaneous form discrimination for retarded children. *Journal of the Experimental Analysis of Behavior, 10*(1), 3-15.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*, 5–22.

Skinner, B. F. (1933a). The rate of establishment of a discrimination. *The Journal of General Psychology, 9*, 302–350. doi: 10.1080/00221309.1933.9920939

Skinner, B. F. (1933b). The abolishment of a discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 19*, 825–828. doi: 10.1073/pnas.19.9.825

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. Oxford, England: Appleton-Century.

Skinner, B. F. (1953). *Science and Human Behavior*. New York: The Free Press.

Skinner, B. F. (1957). *Verbal Behavior*. Cambridge, MA: Prentice-Hall, Inc.

Skinner, B. F. (1974). *About Behaviorism*. New York: Vintage Books.

Skinner, B. F. (1981). Selection by consequences. *Science, 213*, 501–504.

Smeets, P. M., Dymond, S., & Barnes-Holmes, D. (2000). Instructions, stimulus equivalence, and stimulus sorting: Effects of sequential testing arrangements and a default option. *The Psychological Record, 50*, 339–354. Retrieved from: http://opensiuc.lib.siu.edu/tpr/vol350/iss332/338.

Stoddard, L. T., & Sidman, M. (1967). The effects of errors on children's performance on a circle-ellipse discrimination. *Journal of the Experimental Analysis of Behavior, 10*(3), 261-270.

Article 2

The Sorting Test: On the Measurement of Equivalence Class Formation

Sjur Granmo

Faculty of Health Sciences

Department of Behavioral Sciences

Abstract

The purpose of the present experiment is to provide data on the aspects of the concordance between the Matching-to-Sample test and the sorting test. In the present study, 20 college students were divided into two order-controlled groups. Group 1-participants were exposed to an immediate sorting test after training of baseline relations, followed by the administration of a Matching-to-Sample (MTS) test, and another sorting test. Group 2-participants were exposed to an immediate MTS test after training of the baseline relations, followed by a sorting test and a readministration of the MTS test. The results show systematic replication of earlier findings with respect to concordance between the tests, dissociation in some cases, and approached performances (i.e., one or two, of three possible, classes established). However, based on the experimental design of the present experiment, we are able to more effectively discuss the role of equivalence classes in relation to the sorting test outcome. The assumption that one can document equivalence classes by the sorting test alone is rejected with data in the present experiment.

Keywords: sorting, measurement, stimulus equivalence, class formation

The field of stimulus equivalence has been dominated by experiments using matching-to-sample (MTS) procedures to test if the directly trained conditional discriminations (i.e., arbitrary stimulus relations) can be shown to possess the properties of equivalence relations. The MTS-procedure has also been dominantly used in measuring class formation of stimuli that are trained (Fields et al., 2014).

We can illustrate the training and testing of stimulus equivalence research using letters and numerals. For example, three experimenter-defined classes may have been predetermined prior to training and testing. Each of the classes are named by a number; Class 1, Class 2, and Class 3. Letters, on the other hand, indicates the members of the classes and each letter represents a whole stimulus set of three members from the three different classes. That is, the stimulus set A would consist of A1, A2, and A3 (i.e., the number indicating class affiliation). In the literature we can read that A was trained to B, and that means that A1, A2, A3 was all in succession, or randomly, trained to B1, B2, B3.

This can be illustrated in a straightforward example. The participant sits in front of the computer. The stimulus A1 appears as the sample stimulus in the center of the screen. When the participant makes an observing response on the sample stimulus A1, the three comparison stimuli appears in three of the corners of the screen. The comparison stimuli are B1, B2, and B3 (i.e., A is trained to B). To respond correctly on this trial, the participant has to press B1 in the presence of A1. The participant will in such a case receive programmed consequences such as “correct” from the computer program. After an intertrial interval, the participant will be again being presented with a sample stimulus in the center of the screen. This time, the sample stimulus is B2. Upon a response to the sample stimulus the comparison stimuli will appear. The comparison

stimuli are now C1, C2, and C3. A correct response will be C2 upon having B2 as a sample stimulus (i.e., B is trained to C).

In the *linear series* (LS) training structure, the stimulus set A (i.e., A1, A2, A3) are trained to all of the stimuli in stimulus set B (i.e., B1, B2, B3), and all of the stimuli from stimulus set B are trained to all of the stimuli in stimulus set C (i.e., C1, C2, C3). The trained relations are specified with letters (i.e., AB = training all A's to all B's), whereas the trials are specified with letters *and* numbers (i.e., A1B1 = training trial consisting of A1 as sample, and all B stimuli as comparison with B1 indicating the correct choice). Upon correct responding in the training trials, we say that three 3-member classes are established. As for procedural details, we could easily expand the members within the classes, or the number of classes itself, at this point. Let us imagine that we added stimulus set D and stimulus set E to the list.

Now, we have to train the CD relation and the DE relation as well. There is seldom presented a "round" of A1B1, A2B2, A3B3, and then presenting B1C1, B2C2, B3C3, and so on. Common to most experiments are the mixing or randomization of trials presented. For example, a participant may start out with the D3E3 trial, and have A2B2 as the next trial, then C3D3 and so on. When all trials are simultaneously mixed so that AB, BC, CD, DE relations are trained altogether in one block of training trials; the training protocol is called *concurrent*. The LS training structure and the simultaneous concurrent protocol are sometimes applied in combination (e.g., Arntzen et al., 2015; Eilifsen & Arntzen, 2009; Nedelcu, Fields, & Arntzen, 2015), because it has been shown to be useful by its low yields of participants establishing the stimulus classes and thereby also participants that are not responding in accordance with stimulus equivalence (Arntzen et al., 2010; Arntzen & Holth, 1997, 2000; Buffington, Fields, & Adams, 1997; Eilifsen & Arntzen, 2009; Fields, Reeve, Rosen, et al., 1997). Low yields of class formation are preferred

when the purpose is to look at the small differences between outcomes of different parameters. Another training format are the *serialized* training protocol. First, we train AB and BC, then we mix all the possible AB and BC relations. Upon mastery of the AB and BC in a mixed training block, we implement CD relations and mix all the AB, BC, and CD relations, after the mastery criterion is reached. At last, we include the DE relation and after a block of DE training trials are mastered we mix AB, BC, CD, and DE training trials in a block.

The training procedure described here is known as *conditional discrimination* (e.g., Sidman & Stoddard, 1967). The test would normally include the testing of trained relations without reinforcement to document the acquisition of the trained discriminations (i.e., also called *baseline relations*). In equivalence research, however, Sidman (1971) found that not only are the trained relations established, but some tests reveal that also other relations may emerge as a result of the training of certain relations. More specifically, he labeled the emergent relations as equivalence relations and defined them as including properties of *reflexivity*, *symmetry*, and *transitivity* (e.g., Sidman & Tailby, 1982b).

The testing procedure of reflexivity, symmetry, and transitivity is illustrated in the following. Reflexivity is demonstrated if the directly trained stimuli can be related to themselves (Sidman, 1992). For example, if A1 is presented as sample stimulus and A1, A2, and A3 is presented as comparison stimuli, then A1 is the correct response to the comparison stimuli. Symmetry is when sample and comparison is interchangeable (Sidman, 1992). When we have the conditional relations of AB and BC established, the symmetry relation would be if BA and CB emerged. For example, when B1 is presented as sample stimulus and A1, A2, and A3 is presented as comparison stimuli, then A1 is the correct response to the comparison stimuli. Transitivity requires that AB and BC is established as conditional discriminations, and that AC emerges when

tested. For example, when A1 is presented as sample stimulus and C1, C2, and C3 is presented as comparison stimuli, then C1 is the correct response to the comparison stimuli. However, transitivity can emerge in the absence of symmetry, and symmetry can emerge in the absence of transitivity (Sidman, 1992). Therefore, a final test is required to see if the conditional relations possesses both symmetry and transitivity; the equivalence test (e.g., Sidman, 1992). In our example, the CA-relation is tested. For example, if C1 is presented as sample stimulus and A1, A2, and A3 is presented as comparison stimuli, then A1 is the correct response to the comparison stimuli. A positive outcome of the equivalence relation test suggests that the conditional relations are both symmetrical and transitive.

The stimuli used in most equivalence research are *arbitrarily related* to each other (for an exception see Watt, Keenan, Barnes, & Cairns, 1991). If one tree is bigger than another or that two houses are identical to each other, the relation between the stimuli are not very likely to be characterized as arbitrary. However, unless the stimuli have acquired a shared function through a history of reinforcement, they can be called arbitrarily related and each stimulus itself is often *abstract* to the participant. That is, even stimuli that are familiar to a skilled and normally functioning child may be arbitrary to a child that has not learned the relations between them yet. Examples of arbitrary stimulus relations are the ones used in Sidman (1971). The experimenter may say “car” which becomes an auditory stimulus to the participant (A1). A car can be displayed as visual stimuli such as a photograph of a car (i.e., B1) and a text that reads *car* (C1). The same may be arranged for other objects and concepts (e.g., boats, planes, bicycles).

The point of illustrating the arbitrarily related stimuli in Sidman (1971) is that stimuli—once unrelated—requires a training history to be related to each other, but after such training all stimuli occasions the same responses. This is also the case in experiments conducting in

equivalence research today. As in the present experiment, it is often conducted experiments with abstract and arbitrarily related stimuli such as predetermined experimenter-defined relations among Hebrew and Arabic letters from which the participants have no history with. After establishing the baseline relations the responses once made only to two stimulus relations (e.g., A1B1) are now generalized to all stimuli within a class (i.e., A1B1C1D1E1), whereas the same stimuli from each class are discriminated from the other classes (Keller & Schoenfeld, 1950; e.g., members from Class 1 as sample does not occasion the choice of members from Class 2 or 3).

Class formation is the general term used for the establishment of a stimulus class in which the members of the class have a common effect on behavior (Pierce & Cheney, 2008). Equivalence classes contains arbitrary stimulus relations which have become interchangeable to each other as result of contingencies of reinforcement (Sidman, 1994) and the relations within the class are—as the definition indicates—reflexive, symmetrical and transitive. The Matching-To-Sample (MTS) procedure is probably the best way—perhaps even the only way—to test for all the properties of the conditional relations in which define stimulus equivalence. Also, class formation performance after the training of conditional discriminations has been tested dominantly with the MTS-procedure. However, there may be many ways to infer class formation based on test performances. Most of the experiments on equivalence class formation have been concerned with contingencies of reinforcement, training structures, training and testing protocols, trial formats, and the nodal structure of the class (Fields, Arntzen, Nartey, & Eilifsen, 2012). Nevertheless, over the last few decades, other ways have been provided to measure if trained stimulus relations have been established into stimulus classes.

For example, Dymond and Rehfeldt (2001) yielded an overview regarding other ways of measuring class formation in which consisted of response latency/reaction time, verbal reports,

stimulus recall, and stimulus sorting. In particular, Dymond and Rehfeldt reported that sorting tasks was widely applied the research provided from other disciplines than behavior analysis (e.g., Ludvigson & Caul, 1964; Rosch & Mervis, 1977). In behavior analysis, however, only a few studies were reported in using a sorting task as a supplementary measure (Green, 1990; Pilgrim & Galizio, 1996; Smeets, Dymond, & Barnes-Holmes, 2000). Furthermore, Dymond and Rehfeldt (2001) reported that there was (1) close concordance between the subjects' responding in accordance with equivalence and the class formation demonstrated by the sorting test, and (2) a cost-effective measure of the emergence of stimulus relations in which becomes advantageously for the applied researcher.

More recently, however, the inclusions of sorting tests in experiments on equivalence classes have increased slightly (e.g., Arntzen, 2004; Arntzen, Braaten, Lian, & Eilifsen, 2011; Arntzen et al., 2015; Cowley, Green, & Braunling-McMorrow, 1992; Eilifsen & Arntzen, 2009, 2011; Fields et al., 2014; Fields et al., 2012; Fienup & Dixon, 2006; Hove, 2003; Lowe, Horne, Harris, & Randle, 2002; Mackay, Wilkinson, Farrell, & Serna, 2011; Sigurdardottir, Mackay, & Green, 2012).

There seem to be an agreement of the cost-effective utility of the sorting test. Fields et al. (2012) showed concordance between the sorting test and MTS-based tests and suggested sorting test as more time-efficient with respect to their measures of about 2 minutes durations contrasted by the MTS-based test in which lasted about 30 minutes. There is widely argued that there is close concordance between the sorting test and the MTS-test. However, based on some recent findings (e.g., Arntzen et al., 2015; Fields et al., 2014; Fields et al., 2012) the preciseness of the concordance between the two tests seems to need further attention.

A typical MTS emergent relations test can discover the properties of equivalence relations, contrasted by the sorting test, in which only an emergence of class formation as a whole can be seen (e.g., Arntzen et al., 2015; Fields et al., 2014; Fields et al., 2012). Nonetheless, when class formation is established, the important aspect of the formed stimulus classes is that the function of the stimuli within classes are generalized to each member and that they are discriminated from stimuli of other classes (Keller & Schoenfeld, 1950). But do the two tests for class formation always show the same outcome?

In Eilifsen and Arntzen (2009), 20 participants were trained to form three 3-member classes in an LS training structure in simultaneous concurrent protocol. After the baseline relations were established, they were tested for emergent relations in a MTS test and for class formation in a subsequent sorting test. Six participants did not respond in accordance with stimulus equivalence on the test for emergent relations using the MTS format. However, in the sorting test the same participants sorted the stimuli into classes in which matched the experimenter-defined relations between them, thereby demonstrating a dissociation between the two tests. Interestingly, another participant did not sort the stimuli correctly, but responded in accordance with stimulus equivalence on the MTS test. This is still the only demonstration of a dissociation where the MTS test was mastered but not the sorting test. Therefore, Eilifsen and Arntzen (2009) concluded that the MTS test and the sorting test yielded different outcomes in performance.

In Fields et al. (2012), participants were to form three 5-member stimulus classes in a serialized protocol using the LS training structure. They did not find such “pure” dissociations between the tests as reported by Arntzen and Eilifsen (2009). However, there were 15 out 30 participants who neither responded in accordance with equivalence class formation in the MTS

based nor the sorting test. Out of these 15, four participants showed some interesting sorting test data. In the sorting test, three of the four participants responded in accordance with one of the three experimenter-defined classes, whereas the remaining participant responded in accordance with four (instead of five) experimenter-defined members in two of the classes. Fields and colleagues described it as "…a dissociation of class-indicative responding occasioned by the two different tests for class formation" and "…a dissociation between the results of the card sorting and the emergent relations test" (p. 173) respectively. In the cases where at least one, but not all, stimulus classes are sorted in a way that matches the experimenter-defined classes we have nominated as "approached performances". Furthermore, they proposed two implications of these results. The sorting test may either be (1) a more sensitive measure of class formation that emergent relations test, or (2) there may be a delayed emergence of the stimulus classes since none of the classes emerged in the MTS-based emergent relations test, or (3) some combination of the two (Fields et al., 2014; Fields et al., 2012).

Arntzen, Nartey, and Fields (2014) trained 40 university students with serialized LS training structure to form three 5-member classes, and who were subsequently tested for the presence of emergent relations and the maintenance of the classes documented in a sorting test. Out of the 40 participants, 15 responded in accordance with stimulus equivalence in the MTS test for emergent relations, whereas 25 participants did not. Two participants showed dissociation between the two tests. That is, they did not respond in accordance with stimulus equivalence on the MTS test administered after the training of baseline relations, but they sorted the stimuli in a way that matched the experimenter-defined classes. Six participants showed the presence of one or two stimulus classes in the sorting test (i.e., approached performance). Arntzen et al. (2014) concluded that the sorting test showed a general close concordance with the MTS test.

Nartey, Arntzen, and Fields (2014a) also trained 40 university students to form three 5-member classes with a serialized training structure. After testing for emergent relations, they were exposed to the sorting test. As in Arntzen et al. (2014), 15 participants responded in accordance with stimulus equivalence, whereas 25 did not. One participant showed the presence of all three experimenter-defined classes in the sorting test without responding in accordance with stimulus equivalence. Eight participants showed the approached performances, that is, they sorted only one or two experimenter-defined classes correctly. Note that we include those who showed delayed emergence within each MTS test in our summary of these previous experiments. Nartey et al. (2014a) concluded in their study with close concordance between the MTS test and the subsequently administered sorting test.

In yet another study by Nartey, Arntzen, and Fields (2014b), 50 university students were trained in a serialized order and LS structure to form three 5-member classes. In all 14 participants showed performances that met the criterion of stimulus equivalence, whereas 36 participants did not. In this study, there were a total of six participants that did not respond in accordance with stimulus equivalence, but who sorted the classes correctly as they were defined by the experimenter. Four participants showed what we call the approached performance where one or two experimenter-defined classes were documented, but however, not all three. Nartey and colleagues concluded that the two tests showed close concordance.

Fields et al. (2014), as all the studies hitherto, trained three 5-member classes using a Linear Series (LS) training structure where the baseline relations was trained in a serialized order. After establishing the relations in training, the MTS-based emergent class formation test was conducted. After the MTS-based test they conducted a sorting test. Fields et al. (2014) described the findings as close concordance between the MTS-based test for equivalence class formation

and the sorting test for class formation. Even in the case of five participants where trial-by-trial analyses revealed that only one class emerged (i.e., approached performance) during the MTS-based test, the sorting test yielded the same outcome. For these participants, the other two classes were not apparent in any of the tests. Fields and colleagues then concluded that the sorting test is a valid measure of the formed classes. As in the studies described up to this point, Fields et al. (2014) also found that a participant—who did not form classes in the MTS-based test—did indeed formed the classes in the subsequent sorting test. However, as relevant for all the described studies, when testing for class formation in the traditional training-and-testing-order (i.e., a sorting test after the MTS test), the sorting performances that are shown to match the experimenter-defined classes can either be interpreted as (1) outcomes that are measured only by the sorting test, but not the MTS test or (2) an outcome in which represents the delayed emergence of the classes.

In attempt to answer questions about the two possibilities proposed by Fields et al. (2012) and Fields et al. (2014) regarding delayed emergence or test sensitivity, Arntzen et al. (2015) designed an experiment in which 16 students were exposed to a simultaneous concurrent training protocol (i.e., not Serialized) where three 5-member classes was to be formed in an LS training structure. They were exposed to the traditional way of testing, (i.e., emergent relations and then sorting performances). That is, the 16 participants were exposed to (1) a sorting pretest, (2) training on the baseline relations, (3) thinning of consequences in a gradual manner, (4) testing for emergent relations in a MTS test, (5) and at last a sorting posttest.

The three participants that responded in accordance with stimulus equivalence on the MTS test were retrained with a new stimulus set. Thereafter, they were retested with the new stimulus set. In this second phase, however, the students received the sorting test *before* the MTS

test, thereby documenting an immediate emergence of the class formation after training for the first time (but compare with the sorting test of Grimm, 2011). That is, the three participants who responded in accordance with stimulus equivalence in the first phase were exposed to (1) a sorting pretest, (2) training of the baseline relations, (3) gradual thinning of the consequences, (4) a sorting posttest, and at last (5) a MTS test for emergent relations.

Altogether, Arntzen et al. (2015), found that the sorting test showed close concordance with the MTS test. This was shown specifically in the three participants who showed the immediate emergence of the relations in the first MTS test with one stimulus set, while in the second phase of the experiment showed the immediate emergence of the classes in the sorting test with another stimulus set. In the first phase of the experiment (i.e., first stimulus set), the same three participants also showed the maintenance of the emergent relations in a sorting test administered after the MTS test. Two other participants showed concordance between the two tests in Phase 2 of the experiment, but they received training and testing with respect to only one stimulus set (due to delayed emergence within the 2-block MTS in Phase 1). Other examples on concordance were the nine participants who responded incorrectly on both tests.

Regarding the dissociations, two participants showed only approached performances in the sorting test in Phase 1, whereas in Phase 2 of the experimenter, they sorted the stimuli in accordance with the experimenter-defined classes. With these results, Arntzen et al. (2015) demonstrated—for the first time—the ability of the sorting test to document the delayed emergence of the formed stimulus classes after a previously incorrect sorting test. The results also show the dissociation previously documented (Eilifsen & Arntzen, 2009; Arntzen et al., 2014; Nartey et al., 2014a; Nartey et al., 2014b; Fields et al., 2014). However, the two participants who showed the dissociation between the two tests also led to some ambiguous interpretations. That

is, they received the traditional order of training and testing and therefore the authors were unable to answer the questions proposed by Fields et al. (2014) regarding sensitivity vs. delayed emergence.

Another claim by Arntzen et al. (2015) is that one cannot say whether the classes formed correctly in the second-phase sorting were equivalence classes or stimulus classes because the correctly performed sorting test was not followed by a MTS test in which could document the emergent relations that defines stimulus equivalence. They wrote: “The results of prior experiments, however, support the expectation that such a test, if it had been conducted, would have shown criterion level responding, and the delayed emergence of the equivalence classes”. As we have seen in the described experiments also referred to by in this quote (Eilifsen & Arntzen, 2009; Arntzen et al., 2014; Nartey et al., 2014a; Nartey et al., 2014b; Fields et al., 2014), a concordance between the MTS test and the sorting test may not always be the case. In summary, however, Arntzen et al. (2015) demonstrated the following qualities of the sorting test: (1) the immediate emergence of class formation, (2) the delayed emergence of the class formation in the sorting test, (3) the maintenance of the classes in the sorting test after correct responding in the MTS test, (4) the maintenance of the classes in a MTS test after correct sorting performance. They concluded that equivalence classes are relatively independent of the test type used to document them. Regarding what the results of Arntzen et al. (2015) would have been if the participants had no prior history with the contingencies are unclear.

The general purpose of the present experiment was to compare the MTS-based class formation test with the sorting test for class formation. A more specific purpose was to systematically replicate the findings of Arntzen et al. (2014), Arntzen et al. (2015), Eilifsen and Arntzen (2009), Fields et al. (2012), Fields et al. (2014), Nartey et al. (2014a), and Nartey et al.

(2014b) regarding the concordance and any dissociation between the MTS test and the sorting test. Further, the approached performances seen in participants from prior experiments are also assessed. The present experiment is also designed to allow findings regarding the ability of the sorting test to document immediate and delayed emergence of the classes, and the maintenance of the classes.

As illustrated in Table 1, two groups of participants were exposed to the following conditions. Both Group 1 and Group 2 participants were exposed to (1) a sorting pretest, (2) training of the baseline relations, and (3) a gradual thinning of the consequences. Then, Group 1-participants were exposed to (4) a sorting posttest, (5) a MTS test for emergent relations, and at last (6) a second sorting posttest. Whereas the Group 2-participants received the (4) testing for emergent relations in a MTS test, (5) a sorting posttest, and at last (6) a MTS test for emergent relations. The present experiment is designed to control for any order effect from one condition to subsequent conditions. Moreover, we need not be concerned with participants having history with the contingencies from earlier phases, contrasting Arntzen et al. (2015). Regarding the dissociation between the two tests, the present study can provide important findings to the conceptual discussion on whether the classes formed in a sorting test can be termed equivalence classes or stimulus classes (discussed in Arntzen et al., 2015). This was accomplished by allowing participants who sorted correctly after an incorrect MTS test to be retested with a MTS test. In other words, a correct performance in the sorting test was given the opportunity to be acknowledged as having the properties of stimulus equivalence in a subsequent MTS test.

If Group 2 participants performed correctly in the sorting test between two correct MTS tests, such an example would also indicate that the sorting performance was most likely equivalence classes (e.g., Arntzen et al., 2015). However, if the participant does not respond in

accordance with stimulus equivalence on the MTS tests, the correctly performed sorting cannot be termed equivalence classes. If the latter can be demonstrated we cannot trust the performance on a sorting test alone to be recognized as equivalence classes.

Method

Participants and setting

Twenty undergraduate health and social care college students (two males and 18 females) between the ages of 20 and 44 years old participated. The setting was an office room (3 x 4 meters) stripped for all unnecessary items possible helping equipment (e.g., pencils, sheets, phone). The participants were sitting at a desk (1 x 1 meters)—facing the wall—with the computer in front of the participant. The experimenter was sitting in a small room (1 x 3 meters) outside the door to the office. After the experiment was conducted the participants received debriefing in which they were informed about the purposes of the study as well as their performances in the experiment.

Informed Consent

Before the experiment began, each participant was asked to sit down in the office to read the general information about the experiment. Mainly, the participants received information on their role as a participant in an experiment within behavior analysis, and that the experiment would be concerned with their performance in different tasks on a computer. They were assured that no harmful effects were predicted in the experiment. They were also assured anonymity and that they could withdraw from the experiment at any time without suffering any consequences. The participants were informed on the day they were recruited that the experiment would take about 4 hours per participant.

Apparatus

The apparatus was a HP Compaq nc6320 PC with Windows 32-bit operating system. The Processor was a 1.83 GHz Genuine Intel ® with 2 GB memory (RAM). The stimuli used in the present experiment was Arabic and Hebrew letters; the same as those used in Arntzen et al. (2015), Fields et al. (2012), and Fields et al. (2014).

Procedure

The to-be-trained relations were three 5-member classes consisting of the relations AB, BC, CD, and DE. The training structure was Linear Series and was the same as in Fields et al. (2012), Fields et al. (2014), and Arntzen et al. (2015). The MTS format (Simultaneous) and order (Concurrent) was the same as Arntzen et al. (2015; see Introduction for more detailed description of the MTS procedure, the training structure, and the training order). Every task included in the present experiment was presented to the participants on the computer screen. The experiment conducted one training phase and three test phases (see Groups and Conditions).

Training of Baseline Relations. In training and testing trials, the sample stimulus appeared in the center of the screen whereas the comparison stimuli appeared in the corners of the screen and their position was randomized by the software program from trial to trial. An observing response (i.e., a mouse-click) was required on the sample stimulus for producing the appearance of comparison stimuli. In training trials only, a correct response to the comparison resulted in programmed consequences such as “good”, “correct”, etc. When an incorrect response to the comparison was emitted, the programmed consequence “wrong” appeared. The interval in which the programmed consequences appeared was set to 500 ms, whereas the Inter-Trial Interval (ITI)—the interval between the choice of comparison stimuli and the appearance of the next sample stimulus—was set to 1000 ms.

In the training of baseline relations, the participant was presented each relation (A1/B1B2B3, i.e., training AB with Class 1 as correct) five times in a block, each block consisting of 60 trials. The mastery criterion for the training trials was 100 % correct in one block of 60 trials in order to proceed to the test phases. Toward the end of the training, there was programmed a thinning of the consequences to document the maintenance of the baseline relations. The thinning of consequences were administered by the software program in the following steps: 100 % programmed consequences in one block, and then—given 100 % correct in that block—the participant received only 50 % programmed consequences in the next block, and at last 0 % programmed consequences in the last block of the training phase. Throughout both training and test phases, the order in which each relation was presented on the screen was randomized (e.g., if the first trial were A3/B1B2B3, then the next trial could be D1/E1E2E3, and so on).

The instruction to all the participants is presented below. The participants were receiving the same instruction translated into Norwegian:

> A stimulus will appear in the middle of the screen. Click on this by using the computer mouse. Then three other stimuli will then appear. Choose one of these using the computer mouse. If you choose the stimulus we have defined as correct, words like very good, excellent, and so on will appear on the screen. If you press a wrong stimulus, the word "wrong" will appear on the screen. At the bottom of the screen, the number of correct responses you have made will be counted. During some stages of the experiment, the computer will not tell you if your choices are correct or wrong. However, based on what you have learned so far, you can get all of the tasks correct. Please do your best to get everything right. Good luck! (Fields et al., 2012, p. 168)

Testing for Emergent Relations and Class Formation. The testing of emergent relations (i.e., symmetry, transitivity, and equivalence) was administered with a MTS test. The MTS test was organized as the MTS training of baseline relations except that the sequences of trials for testing emergent relations were mixed throughout two blocks of 180 trials each (i.e., not 60 trials). The 180 trials were presented because each of the possible relations was presented three times (i.e., not five times per relation). The possible relations were either symmetrical (i.e., BA, CB, DC, ED), transitive (i.e., AC, AD, AE, BD, BE, and CE), or equivalence relations (i.e., CA, DA, EA, DB, EB, and EC). Therefore, one block of 180 trials consisted of 36 baseline relations, 36 symmetrical relations, 54 transitive relations and 54 equivalence relations. The criterion that would document stimulus equivalence was set to 95 % correct of the trials in each of the two blocks (i.e., marked in bold face in Table 2).

Sorting test. In the sorting test, one of the stimuli was displayed on the screen. Behind the first stimulus the remaining stimuli laid in a way that could be recognized as a stack of cards. The participants were told that if they dragged the stimulus to the side, using the computer mouse, a second stimulus would appear behind where the first stimulus originally had laid. They were then told that this would repeat itself until there were no more stimuli left (i.e., a total 15 stimuli). The participants were then told to "drag all the stimuli apart from each other so that all stimuli lays visible in front of you" and "put them together the way you think is correct, but make sure that all stimuli are visible" (i.e., to prevent that participants was placing the stimuli on top of each other). Finally, the participants were told to call on the experimenter when they were finished. The experimenter took a screenshot of the way the participant had sorted the stimuli. The screenshot picture was automatically recorded and saved in a document folder on the computer.

After the experiment was conducted, two observers successively assessed the stimulus combinations from the sorting-test screenshot in an Inter Observer Agreement (IOA). Both observers scored the sorting data with respect to how many classes one could identify based on the participant's sorting performance. Data from the pre class-formation sorting test were hardly ever clustered together in groups across participants. Therefore, these data were categorized as either correct (i.e., problematic to validity) or as incorrect (i.e., not problematic). None of the participants had pre class-formation sorting performance in which was problematic; all performances were incorrect. The two observers scored all the remaining participants with 100 % agreement regarding whether the participant had sorted the stimuli in accordance with one or more experimenter-defined classes. In cases where the participants sorted the stimuli in ways that were incorrectly regardless of how the observer interpreted the image, it was considered to be irrelevant if disagreement between the observers occurred (e.g., the case illustrated in Figure 7).

For each participant, independent of group assignment, the first task was the sorting pretest (PreSRT-1). The PreSRT-1 would determine whether the participants had any history with the relations among the stimuli in which was defined by the experimenter.

Dependent on group assignment, the sorting test was administered either (1) immediately after the establishment of the directly trained baseline relations (i.e., allowing the demonstration of immediate emergence in a sorting test), or (2) after a 2-block MTS emergent relations test.

Groups and Conditions. The participants were assigned randomly into two groups, each containing 10 participants. Participants assigned to Group 1 were exposed to the following order of training and testing: (1) pre class-formation sorting test (PreSRT-1), (2) training baseline relations until all trials in a block achieves 100 % correct, (3) thinning of consequences, (4) a post

class-formation sorting test (SRT-1), (5) two blocks of MTS emergent relations test (each block consisting of 180 trials), and finally (6) a new post class-formation sorting test (SRT-2).

Participants assigned to Group 2 were exposed to the following order of training and testing: 1) pre class-formation sorting test (PreSRT-1), (2) training baseline relations until all trials in a block achieves 100 % correct, (3) thinning of consequences, (4) a 2-block MTS emergent relations test, (5) post class-formation sorting test (SRT-1), and finally (6) a 2-block MTS emergent relations test.

The term "post class-formation sorting test" (SRT-1/SRT-2) should be understood as procedural language, not as a process (i.e., the participants may, of course, not perform in accordance with class formation at all).

Dropouts and Remarks

Two participants were withdrawn in the present experiment. The participants were told on the recruitment day that the experiment would last no longer than 4 hours. Participant 15116 exceeded this 4-hour time limit during training and did not wish to complete the participation of the experiment. Although instructed to sort the stimuli in a way that would make all stimuli visible for inspection, Participant 15112 sorted the stimuli in a way that made it impossible to examine the stimuli by the screen shot (i.e., put the stimuli in stacks). Two other participants, namely 15121 and 15122, replaced these two participants. Notice that even and odd numbers in the Participant Number (PN) do not necessarily mean that participants belonged to Group 1 and 2, respectively.

A remark should be made in the case of Participant 15101 who did not start the experiment with the pre class-formation sorting test, in a way similar to all the other participants.

Instead, she started off with the training of baseline relations immediately after signing the informed consent. The pre class-formation sorting test was introduced about 5 minutes into the training of baseline relations. However, even when experiencing 5 minutes of training before the pre class-formation sorting test, the participant's performance do not suggest that this mistake brought with it any confounding effects (see Table 2).

Results

In Table 2, the performances of each participant are presented from the pre class-formation sorting test (i.e., Pre-SRT), the training of baseline relations (i.e., TBR), and post class-formation sorting tests (i.e., SRT-1 and SRT-2) as well as MTS tests (i.e., MTS 1a + 1b, and 2a + 2b). In sorting tests, the participant sorted the stimuli into clusters. Each cluster represents a 3-digit string (e.g., 221) shown beneath all sorting tests (SRT) in Table 2. The first number in the string represents how many experimenter-defined Class 1-stimuli that was sorted in the participant-defined cluster. The same was true for the second and third number in the string, which represented the experimenter-defined Class 2-stimuli and Class-3 stimuli in that particular cluster, respectively.

For example, during the pre class-formation sorting test (i.e., Pre-SRT), participant 15108 sorted the stimuli in a 301 outcome in the first of her clusters. Such performance means that three stimuli from Class 1 was represented, zero stimuli from Class 2, and one stimulus from Class 3. In participant 15101's 500 050 005-performance in the first sorting test after training shows that five members of Class 1, five members of Class 2, and five members of Class 3 were represented in the three separate clusters. Such performance indicates that all three experimenter-defined 5-member classes corresponded to the participant-defined classes by which we called stimulus class formation.

Pre Class-Formation Sorting Test

The results from the sorting test administered prior to the training baseline relations are presented as Pre-SRT in Table 2. Participants from each groups generated everything in between three and six clusters of stimuli. None of the clusters contained the stimuli that matched the experimenter-defined classes in this phase of the experiment. Therefore, we documented that none of the participants had previous history with any of the relations to-be-trained.

Training of Baseline Relations and Testing of Emergent Relations

Each of the participant's performance is presented in each row in Table 2 read from the Participant Number (PN) in the left column of the table. Between-group analyses are not considered relevant to the present experiment. The separation of participants in two groups had no other purpose than to control for any possible order effect from the conditions that the participants were exposed to. The acquisition of the baseline relations required different amounts of training trials, but with a mean of 1020 trials, ranging from 540–2100 trials. After the training of baseline relations, the phases to which each group was exposed to were individually designed.

Immediate formation of the classes in the sorting test.

Participants assigned to Group 1 were exposed to a sorting test immediately after the training of baseline relations. Five participants (15101, 15103, 15113, 15117, and 15119) showed the immediate formation of the classes in the first post class-formation sorting test (SRT-1 in Table 2). Four of these participants showed the maintenance of the classes in the second administered sorting test (SRT-2 in Table 2). The remaining participants (15105, 15107, 15111, 15109, and 15115) neither formed all the experimenter-defined classes in the first sorting test nor showed the maintenance of all the classes in the second sorting test (SRT-2).

Matching-to-sample performances and concordance between the tests.

Three participants from Group 1 (15101, 15103, and 15119) and three participants from Group 2 (15102, 15118, and 15122) responded in accordance with stimulus equivalence in the MTS test for emergent relation (MTS 1a + 1b in Table 2). The percentage correct ranged from 95–100 %. The three participants from Group 1 also responded in accordance with the experimenter-defined classes in the first post class-formation sorting test (SRT-1). After the MTS test (1a + 1b), they also sorted the stimuli in participant-defined classes that matched the experimenter-defined classes in the second sorting test (SRT-2), thereby demonstrating concordance between the two tests with respect to correct performances. The three participants from Group 2 who responded in accordance with stimulus equivalence in the first MTS test (1a + 1b) did also show the maintenance of the classes in the second MTS test (2a + 2b). The same two participants also sorted the stimuli in experimenter-defined classes in the post class-formation sorting test (SRT-1), and can therefore be added to the cases of concordance between the two tests with respect to correct performances.

Thirteen participants did not respond in accordance with stimulus equivalence in the MTS tests. Their performances ranged from 33–88 % correct. However, nine of the 13 participants who did not meet the criterion for stimulus equivalence showed other interesting performances across the sorting tests and the subsequent MTS test (see "Dissociation between the tests and delayed emergence" and "Approached performances"). The remaining four participants (15109, 15110, 15114, and 15115) did neither respond correctly on the MTS tests nor any of the sorting tests.

Dissociation between tests and delayed emergence.

Two of the participants from Group 1 (15113 and 15117) responded in accordance with the experimenter-defined classes in the sorting test, but did not show the maintenance of the classes in the subsequent MTS test (1a + 1b). Interestingly, Participant 15117 did not show maintenance of the classes in the second administered sorting test, that is, she responded correctly on the first post class formation sorting test (SRT-1), but not on the second post class-formation sorting test (SRT-1). Three of the participants from Group 2 (15104, 15120, and 15121) did not respond in accordance with stimulus equivalence in the MTS test (1a + 1b), but did in fact sort the stimuli in accordance with the experimenter-defined classes in the subsequent sorting test. Two of these three participants (15104 and 15120) from Group 2 did not show the maintenance of the previously demonstrated stimulus classes from the sorting test in the last MTS test (2a + 2b). In contrast, participant 15121 responded in accordance with stimulus equivalence in the second MTS test for emergent relations (2a + 2b), even though she had relatively low scores in the first test of emergent relations (1a + 1b), thereby demonstrated a case of delayed emergence.

Approached performances.

Three participants from Group 1 (15105, 15107, and 15111) showed sorting performances we nominated as "approached performance" (first demonstrated by Fields et al., 2012). That is, one or two of the experimenter-defined classes were documented by the sorting test. Note that such performances are not considered as a performance that meets the criterion of established relations among stimuli within the classes. They are considered as established relations only within that particular class and would be considered as a failed test (e.g., when evaluating the concordance between the tests).

For participant 15105, all experimenter-defined members belonging to Class 1 and 3 corresponded to the participant-defined classes. However, we saw a split between the C2-

stimulus and the D2-stimulus, so that A2B2C2 laid in one cluster and D2E2 laid in another cluster.

For participant 15107 and 15111, the stimuli were sorted into four and five clusters, respectively. For these participants, only one of the clusters contained all the experimenter-defined members of Class 2. Participant 15107 improved her performance in the second sorting test (SRT-2) by sorting the experimenter-defined cluster of Class 2-members, and in addition producing a cluster in which contained all five stimuli from Class 3, but with the inclusion of two stimuli from Class 1. In contrast, participant 15111 did not show the maintenance of the previous "established" class in the second sorting test (SRT-2). However, based on the sorting data in Table 2, one can easily and relatively quick see that Class 1 and 2 are still fairly representing with four experimenter-defined stimuli in each group, but cannot of course, be regarded as fully established stimulus classes. Two participants from Group 2 (15106 and 15108) also sorted the stimuli in accordance with only one of the three experimenter-defined classes.

However, participant 15108 showed the five members of Class 2 in one cluster, but the cluster was in addition, accompanied by a stimulus from Class 1 (i.e., the C1-stimulus, see Figure 1). A trial-by-trial error analysis revealed that many stimulus relations that were linked to the accompanying stimulus was also responded to incorrectly in the second MTS test (see Figure 2 and 3 for the trial-by-trial error analysis). Specifically, when presented with B2 as sample stimulus, the participant responded to C1 on every occasion. Such performances were also the case when B1 was the sample and the experimenter-defined correct comparison stimulus to-be-chosen was C1, but the participant chose the C2-stimulus each time instead. At last, we saw that C1 samples did not occasion the selection of D1, but instead, the participant chose D2 on every occasion.

For participant 15105, 15106, and 15107, the sorting test was the only test that would demonstrate the performance of one or two classes but not all three. That is, the trial-by-trial analysis based on the MTS test did not show any consistently incorrect and participant-defined responding as seen with participant 15108 (cf. Participant 4113, 4134, and 4141 in Fields et al., 2012).

Discussion

The main purpose of the present experiment was to replicate the findings from prior experiments (Arntzen et al., 2014; Arntzen et al., 2015; Eilifsen & Arntzen, 2009; Fields et al., 2014; Fields et al., 2012; Nartey et al., 2014a, 2014b). That is, we designed an order-controlled experiment that could show the concordance between the MTS test and the sorting test, but also any dissociation between the two tests. Furthermore, we sought to replicate approached performances in which the participants sort some, but not all, experimenter-defined classes. The design also permitted possible findings relating to the maintenance and the delayed emergence of the classes by administering a MTS test or a sorting test (depending on group) after the previous tests (see Table 1).

Replications of Prior Experiments

Regarding the concordance shown by prior experiments, the present experiment documented that 15 out of the 20 participants showed concordance between the tests. That is, in cases where either both of the tests, or none of the tests, yielded sufficient performance relative to the criterion set by the experimenter (95 % in the MTS test; 100 % in the sorting test), the performance was regarded as concordance. Six participants, three in each group, met the criterion for both the sorting test and the MTS test. Nine participants, five from Group 1 and four from

Group 2, did not respond correctly on either of the tests (for participant numbers and a more detailed presentation, see Table 2).

Of these nine cases of concordance, five participants showed the approached performance. That is, three participants from Group 1 (15105, 15107, and 15111) and two participants from Group 2 (15106 and 15108) sorted the stimuli in one or two experimenter-defined classes, but leaving the remaining stimuli to be sorted in participant-defined classes only. Trial-by-trial analyses (exemplified for Participant 15108 in Figure 2 and 3) found that Participant 15105, 15106, 15107, and 15111 did not perform in consistent participant-defined performance in the MTS tests, as they did the sorting test. That is, these participant-defined classes were only made visible in the sorting test. The purpose of the trial-by-trial analysis was to see if the results of a sorting test could be regarded as a predictor on a possible consistent incorrect participant-defined performance in the MTS test. Participant 15108 was the only case in which the participant-defined classes demonstrated in the sorting test also were present in the MTS trials. Such detailed concordance between the two tests is thereby replicated from the study of Fields et al. (2014). For the four other participants who showed approached performances, the fact that one or two experimenter-defined classes were documented in a sorting test but not in the MTS test for emergent relations "constitutes a dissociation of the class-indicative responding occasioned by the two different tests for class formation" (Fields et al., 2012, p. 173). In addition, cases of somewhat pure dissociation were seen in five other participants in the present experiment (see next paragraph). Thus, these results suggest that one should not depend on the sorting test alone to predict participant-defined performances in a MTS test.

For the purpose to document any dissociation between the two tests, such performances are seen in two participants in Group 1 (15113 and 15117), and in three participants in Group 2

(15104, 15120, and 15121). For these participants, the dissociation consists of a sorting test that matches the experimenter-defined classes, whereas in the MTS test, the same experimenter-defined classes are not responded to in accordance with stimulus equivalence.

The strength of the present experiment was—at least—to demonstrate that the phenomena of concordance, dissociation, and approached performances could be documented in both groups. In addition, our findings show very little variation with respect to the number of cases in which demonstrated the phenomena in each group, and its validity are considered strengthened by such balance in data across groups.

Performances of Novel Phenomena

Some findings from prior experiments were not seen in the present experiment. First, the delayed emergence demonstrated in a sorting test, in which was reported by Arntzen et al. (2015) was not demonstrated. The closest case of delayed emergence shown by the sorting test was for Participant 15107, who showed maintenance of one class in the second sorting test (i.e., Class 2), but now with the derivation of an additional full cluster of members from another class (i.e., Class 3). However, Class 3 stimuli in the second sorting test were accompanied by two stimuli from the remaining class (i.e., Class 1). Therefore, although we saw one cluster that contained all members of one of the classes, one cannot count such performance as class-consistent responding with respect to the lack of discrimination between the classes (Keller & Schoenfeld, 1950).

The only case of delayed emergence in our experiment was documented by Participant 15121, where the delayed emergence were documented in the MTS test, not in the sorting test. However, such a finding is in itself interesting. This participant demonstrated that the classes were established in the sorting test administered after the first incorrect MTS test (i.e., 1a + 1b) and before the second MTS test (2a + 2b). When the participant showed delayed emergence of

the classes in the second MTS test, she also demonstrated maintenance of the classes shown in the sorting test, but now *with* the emergent relations that was not shown in the first MTS test. This finding stresses the importance of treating the classes (i.e., shown in the sorting test) and the emergent relations (i.e., shown in the MTS test) as two different conceptual entities.

Although the present experiment did not document the delayed emergence in the sorting test—quite contrary—it documented a reduction of classes over subsequent testing in some cases. Participant 15117 sorted the stimuli in experimenter-defined classes in the first sorting test, but in the second sorting test, no experimenter-defined classes were documented. Also in the first sorting test, Participant 15105 showed the presence of all experimenter-defined members of two classes. However, in the second sorting test, only one of the two classes was maintained. For Participant 15111, in the first post class-formation sorting test (SRT-1), all stimuli from Class 2 were sorted into a separate cluster that matched the experimenter-defined class, but the other stimuli belonging to Classes 1 and 3 were not. In the second test, Class 2 was not fully partitioned from the rest (i.e., one member was sorted to another cluster) and none of the experimenter-defined classes were demonstrated.

Therefore, the sorting test is shown to document the reduction of classes as well as delayed emergence of the classes shown in Arntzen et al. (2015).

The finding of Eilifsen and Arntzen (2009), in which one participant showed dissociation where the MTS test was performed in accordance with stimulus equivalence whereas the sorting test documented none of the experimenter-defined classes, was not replicated in the present experiment. Although the phenomenon has been proven to be a possible dissociation of the two tests, replications are lacking in all of the other experiments discussed in the present experiment as well.

Differences in Parameters

There are some differences between the present study and the studies replicated, as well as between all of the studies separately (Arntzen et al., 2014; Arntzen et al., 2015; Eilifsen & Arntzen, 2009; Fields et al., 2014; Nartey et al., 2014a, 2014b). Such tiny differences between the studies in combination with small differences in the findings should be a good example of what Sidman (1960) described as *systematic replications*.

With respect to the number of participants, the present study used 20 participants, in which was considered enough to successfully demonstrate replications in each group, achieve control of order between groups, and show findings that answered the research questions of interest. In the prior experiments, the number of participants has ranged from 16 participants in Arntzen et al. (2015) to 50 participants in Fields et al. (2014). Therefore, the different findings are now demonstrated in an increasing number of participants.

The LS training structure has been applied in all of the experiments presumably due to its ability to yield only a small amount of participants responding in accordance with stimulus equivalence; thereby giving the opportunity to look more closely at the differences between the conditions of interest.

With respect to the training protocols of either serialized or concurrent, Eilifsen and Arntzen (2009) used a concurrent protocol. However, the experiments that followed (i.e., Arntzen et al., 2014; Fields et al., 2014; Fields et al., 2012; Nartey et al., 2014a, 2014b) all used the serialized protocol and their findings are characterized by concordance between the sorting test and the MTS test, contrasting to the experiment of Eilifsen and Arntzen (2009). Arntzen et al. (2015) reintroduced of the concurrent protocol. This time, the concordance was actually replicated from the prior experiments of Arntzen et al. (2014), Fields et al. (2012), Fields et al.

(2014), Nartey et al. (2014a, 2014b). With respect to concordance, dissociations, and approached performance, the present experiment also demonstrates replications of all phenomena with the use of the concurrent protocol.

The mastery criterion for the training of baseline relations has also varied from experiment to experiment. The same is true for the criterion regarding stimulus equivalence in the MTS test. In Eilifsen and Arntzen (2009), the training criterion was set to 89 %, whereas the criterion for responding in accordance stimulus equivalence was set to 94,4 %. Arntzen et al. (2014), Fields et al. (2012), and Nartey et al. (2014a, 2014b) required that the participants performed 90 % of the training trials in a block correctly before proceeding to the thinning of consequences phase (i.e., maintenance of baseline relations). They also required the participants to perform 90 % correct on the MTS test for stimulus equivalence to pass the test. In Fields et al. (2014), the mastery criterion of training was set to 100 % in training trials correct before proceeding, whereas the criterion for demonstrating stimulus equivalence was 90 %. Arntzen et al. (2015) set a mastery criterion of 90 % for demonstrating the mastery of trained baseline relations, and the test criterion was set to 95 %. Now, the remaining combination was applied by the present experiment with the training criterion of 100 % and the stimulus equivalence test criterion of 95 % correct in a block. Therefore, most of the phenomena of interest are demonstrated across different mastery criteria in the training of baseline relation trials.

Another difference between the studies is the instruction that was given to the participants after signing the informed consent and before starting the pre class-formation sorting test (Pre-SRT in Table 1 and 2), but also before each post class-formation sorting test. In Eilifsen and Arntzen (2009, p. 192), the participants were "…told to categorize…" the stimuli. Fields et al. (2012, p. 168) and Fields et al. (2014, p. 489) used the same instruction: "Please put them into

groups and call me when you have completed the task". Arntzen et al. (2014, p. 352), Nartey et al. (2014a, p. 779) and Nartey et al. (2014b, p. 22) told the participants to "put them into groups". Arntzen et al. (2015, p. 5), however, changed the instruction to "Put these into groups as you feel like", but the participants also received the following instruction:

> …were informed that to see the next stimulus in the deck, the top card had to be moved to a different location on the screen that was close to other related stimuli so that they formed a cluster that was separated from other clusters that contained stimuli from different sets.

In the present study, we presented the participants with yet another instruction, namely: "Drag all the stimuli apart from each other so that all stimuli lays visible in front of you" and "put them together the way you think is correct, but make sure that all stimuli are visible".

Summarized, each study reveals more or less the same outcomes in spite of the small changes in the instructions presented to the participants across the sorting tests.

Contributions from the Present Experiment

Although the present experiment have demonstrated concordance, dissociation, and approached performances in both groups, along with examples on delayed and maintained classes, the present experiment also contributes to new questions and conceptual discussions regarding the sorting test as a valid measure of class formation in stimulus equivalence research.

Arntzen et al. (2015) demonstrated the immediate emergence of the classes in a sorting test by administering the sorting test as the first test after training. However, by reintroducing conditions of training and testing to the same participants who received the conditions on beforehand, they may have confounded the interpretation of the participant's subsequent

performance. In other words, having history with the contingencies may be a potential confounding variable. The present experiment showed that the sorting test could document immediate emergence without using experienced participants, and the findings supports Arntzen and colleagues' conclusion.

Fields et al. (2014) raised some questions on whether the classes of the sorting test was an example of delayed emergence or that the sorting test shows other outcomes than the MTS test. By controlling for order effects, the present experiment shows that the previous findings (Arntzen et al., 2014; Arntzen et al., 2015; Eilifsen & Arntzen, 2009; Fields et al., 2014; Nartey et al., 2014a, 2014b) can—at least—be interpreted as different outcomes between the sorting test and the MTS test.

The fact that a dissociation between the two tests is demonstrated in each of the groups in the present experiment support the interpretation that dissociation can be the outcome regardless of the order in which the tests are administered. Therefore, when Fields et al. (2014) did not know whether the correct responding of their participants was a case of delayed emergence or that the sorting test detected something other than the MTS test, the present experiment suggests that both assumptions could be the case, in Fields and colleagues' experiment that is. Contrary, in the present experiment, three participants (15104, 15120, and 15121) show the same pattern of responding as the participants of Fields and colleagues' (2014) experiment. If our experiment had ended after the sorting test, we would have to ask the same question as Fields and colleagues. However, our experiment was designed to better detect delayed emergence, and in the subsequent MTS test (2a + 2b) we can see that two of these three participants did not maintain the classes that was documented in the sorting test. Therefore, for these two participants, the classes documented by the sorting test were not likely to be an example of delayed emergence; there was

dissociation between the tests. These findings from Group 2 is also replicated in two participants in Group 1 (15113 and 15117); thereby controlled for any order effect.

Based on these findings, we are left with the assumption that the sorting test most likely shows different outcomes than the MTS test. In our experiment, the second MTS test detected delayed emergence. However, a replication of the performance of Participant 15121 is not possible to achieve in Group 1, of course, with the order of conditions administered by the present experiment. Therefore, an ultimate control for delayed emergence—if possible—is beyond the scope of the present experiment.

Note that the traditional design (i.e., MTS after training and a subsequent sorting test) is not impaired by the findings of the present experiment. The main contribution on the issue of dissociation from our study is to show that the classes can better be evaluated for delayed emergence or maintenance with a subsequent MTS test administered after the sorting test. Subsequent MTS tests can also give indications that the classes which are documented does in fact have the properties of stimulus equivalence (e.g., Arntzen et al., 2015). Such an improvement of the design is considered by the authors of the present study as an important contribution to the discussion on whether there is dissociation between the sorting test and the MTS test.

Equivalence Class Formation vs. Trained Relations

As emphasized in the introduction, the MTS test actually test for responding in accordance with the emergent relations that defines stimulus equivalence. It has been proposed that if a subsequent MTS test is administered after the post class-formation sorting test—and the responding is in accordance stimulus equivalence—one can demonstrate whether the classes partitioned in the sorting test are equivalence classes (e.g., Arntzen et al., 2015). Regarding their sorting data, Arntzen et al. (2015, p. 8) wrote: “While this sorting test documented the formation

of stimulus classes, the classes may not have been equivalence classes because only an indeterminate subset of all of the derived relations was presented in the sorting test". Arntzen and colleagues (2015) at last concluded with the following:

> …mastery in the first MTS tests documented the presence of all relations in the equivalence classes, along with the maintenance of the previously emergent classes. Most likely, then, the classes that emerged in the sorting test [afterwards] were actually equivalence classes. These results support the view that a sorting test can document equivalence class formation. (p. 8)

Now, when looking at the dissociation between the sorting test and the MTS test, those participants show similar sorting performances to each other, along with not respond in accordance with stimulus equivalence. More specifically, the participants that sorted the stimuli correctly, but did not master the test for stimulus equivalence, actually sorted the stimuli in the way the relations had been trained (e.g., Figure 4). When the same participants do not show the emergence of stimulus equivalence in the subsequent MTS test (2a + 2b), it leads us to expect that the performances in the previously administered sorting test (SRT-1) was not equivalence classes. In contrast to the last sentence of the quote from Arntzen et al. (2015), this leaves us no other choice than to dismiss such a view on the sorting test, especially if the sorting test is the only test that is administered in a study.

There are additional findings from the present experiment that would support our view. First, when comparing the performance between the participants that did not respond in accordance with stimulus equivalence and with those who actually did so, the sorting tests can be shown to be exactly the same (e.g., upper panel of Figure 5), or very much alike (e.g., Figure 6). Because the performance can be the same in spite of any performance in accordance with

stimulus equivalence, or not, in either of the two MTS tests, one cannot know for certain if the classes—based on the sorting test alone—can be termed equivalence classes. Based on the present experiment, therefore, we suggests that if one shall term the stimulus classes documented in the sorting test as equivalence classes, one will have to demonstrate the emergent relations in subsequent MTS tests for symmetry, transitivity, and equivalence relations. Even when administering such subsequent tests, we are still uncertain that the MTS test and the sorting test will test for the same outcomes. We shall leave the readers with some speculations based on the findings from in the present experiment.

As mentioned already, the sorting performances—in those that mastered the sorting test only—were, in some cases, formed in the exact same topography as the stimulus relations that were trained (e.g., Figure 4). This led us to speculate if the stimulus relations that were documented in the sorting test were not derived, but rather should be regarded as trained relations. If this assumption is correct, one can ask whether a subsequent MTS test—as recommended by Arntzen et al. (2015)—actually would confirm that the stimulus classes documented in a sorting test, is equivalence classes. More research seems to needed on this issue.

Another speculation based on our findings may suggest that the sorting test actually could have helped the participants in subsequent tests. The performance in Participant 15121 shows us that the only procedural event happening between the first incorrect MTS test (1a + 1b) and the second correct MTS test (2a + 2b) was the correctly completed sorting test (SRT-1). This led us to assume the sorting test actually could have helped the participant in responding in accordance with stimulus equivalence on the second MTS test. But how could the sorting test have helped the participant's performance?

In the sorting test, the participants are presented with simultaneous discrimination in which there is "…possible for the participants to scan back and forth between the different stimuli" (Eilifsen & Arntzen, 2009, p. 199). Also, the discrimination analysis by Saunders and Green (1999) on different training structures in stimulus equivalence research suggests that simultaneous discriminations (e.g., sorting test) are more easily learned than more successive discriminations (e.g., MTS test) and also that a simultaneous discrimination not necessarily leads to successive discrimination, in the way successive discrimination may lead to simultaneous discrimination.

Following this rationale, recall Participant 15121, who showed a remarkable increase in correct responses in the second MTS test (i.e., delayed emergence). Could the sorting test, with its simultaneous discrimination of the directly trained relations (i.e., already established baseline relations) have contributed to the emergence seen in the second MTS test? We believe so. That is, the participant could sort "sample" A1 to "comparison" B1 on the computer screen among all the other stimuli, and in the same manner with A2 and A3 to B2 and B3, respectively, and so on with B to C, C to D, and D to E for Classes 1, 2 and 3. Thereby, the participant would have performed in a successive discrimination fashion. After such a performance—only based on the established baseline relations—the participant is looking at a "picture" of all the relations linked together with each other, and at the same time forming clusters of stimuli. Now, the participant need only *see* which of the E stimuli that belongs to each of the correspondent D-, C-, B-, and A-stimuli and so on with the rest of the relations that are not directly trained. It seems to us that the participant is very likely to be influenced by such a display at this point of the experiment (e.g., with possible reflexive conditioned motivating operations [CMO-R; Michael, 1993] resulting from the first MTS-test's extinction trials).

However, it is not ruled out by the present experiment whether the sorting test was the variable producing the delayed emergence in Participant 15121. Although the sorting test was the only event happening between the two tests, it could be a possibility that a delayed emergence occurred in the second MTS test and thereby was an example of the repeated testing (Shadish, Cook, & Campbell, 2002). However, an interesting feature of Participant 15121's data seen in Table 2 is the increase in correct performance from MTS 1b to MST 2a. If the participant had shown a gradual increase in correct responding throughout MTS 1b and still gradually increasing in MTS 2a, one would—to a greater degree at least—consider the improvement in performance to be related to the possible confounding variable of repeated testing. In Participant 15121, however, 62 % correct responses in MTS 1b can be compared to 97 % correct responses in MTS 2a. In the raw data material, we see that 55 and 56 correct responses in MTS 1b was distributed to the first half and second half of the test block, respectively. In MTS 2a, the pattern is five errors in the first half and one error in the second half. The five errors in the first half were related to different stimulus combinations. Trial-by-trial error analysis of the remaining tests showed no consistent participant-defined responding.

Future experiments can contribute to the understanding with the following design: After participants show incorrect performance on a MTS test for emergent relations, one can administer two consecutive conditions of MTS tests for one group while another group could receive the sorting test followed by a MTS test. If participants in the latter group were generally responding more correctly on the final MTS test for the emergent relations, this could further indicate that the sorting test has a great impact on subsequent performance.

Furthermore, it is not an inevitable consequence for the sorting test to produce delayed emergence. The present experiment showed two examples (Participants 15104 and 15120) of

correct performance in a sorting test, from which it did not result in correct performance on the subsequent MTS test. Also, we saw nine examples of participants that had not sorted the stimuli correctly although the training was completed with 100 % correct.

However, we treat the possibility of such a "helping picture" as a plausible occurring event resulting from a correctly performed sorting test. In fact, the visual stimulus resulting from the responding in accordance with baseline relations seem likely to have influence the participant's performance in much of the same way as *precurrent behavior* produces discriminative stimuli for further effective action (Skinner, 1969). It should be noted that in the trial-by-trial error analysis, neither of the participants showing correct sorting-only performance showed consistently incorrect participant-defined responding in the MTS trials.

Implications and Further Research

The sorting test has been recommended as a quick and easy-to-administer test for class formation by many studies (e.g., Arntzen et al., 2015; Dymond & Rehfeldt, 2001; Fields et al., 2014; Fields et al., 2012). Based on the present discussion, we have no objections to the use of the sorting test in applied settings. In applied settings, we need not be very concerned with demonstrating that the sorting test alone yield classes that are, in fact, equivalence classes or just trained relations. However, the issue could be of interest if we try to improve reading comprehension as in Sidman (1971), but in such cases, the sorting test could be accompanied with a MTS tests, and training can even be repeated until the participant responded correctly on any given test. That is, even if the classes are suspected to consist of only trained relations, one can easily test—and even train—the remaining relations. Indeed, we are encouraging the use of sorting tests in such situations. The sorting test seems to have the ability to quickly detect the missing relation between the stimulus relations to-be-trained (see Figure 1). Then, after

identifying the “missing link”, one can intervene and train those specific relations over and over again.

For applied implications, further research should demonstrate such performances with the sorting tests. The MTS-training and sorting tests could be demonstrated in various educational arenas (e.g., Sidman, 1994), and the focus of the studies could be to test our expectations about the easily detectable missing relations between the stimuli. Then, it can be demonstrated that the sorting test performance can be improved by intervening on that particular relation. The training structure of such applied studies should also—if not exclusively—use One-To-Many (OTM) or Many-To-One (MTO) training structures, since these structures have been demonstrated to yield a higher probability of performances in accordance with stimulus equivalence (Arntzen & Holth, 1997). The advantages of a less time-consuming sorting test in which is already reported (e.g., Arntzen et al., 2015; Dymond & Rehfeldt, 2001; Fields et al., 2014; Fields et al., 2012), are supported by the present experiment. The sorting test took about 2–3 minutes to complete, whereas the MTS test took about 20 minutes to complete. Such practically advantages should be preferred in applied settings.

As for the further research with experiments on the mechanisms discussed in the present study, we call for experiments that elaborate on the issue of simultaneous and successive discrimination. For example, one group can receive simultaneous discrimination training of the baseline relations. Another group can receive the traditional successive discrimination of the baseline relations (i.e., MTS). Then one can expose each group for order-controlled tests—as in the present experiment—and the results should show something interesting about the role of successive versus simultaneous discrimination in establishing the and maintaining class formation.

Also, additional MTS tests after the sorting tests could be implemented by future research, so that the MTS test had the opportunity to detect even more delayed emergence. Regarding the instructions given, further research can test various kinds of instructions and document the effect of such. To make the present experiment's procedure more reliable, one can better instruct the participants by stressing to "make sure that the clusters you have decided to group are separated from the other clusters", thereby preventing disagreement on an IOA.

References

Arntzen, E. (2004). Probability of equivalence formation: Familiar stimuli and training structure. *The Psychological Record, 54*, 275–291.

Arntzen, E., Braaten, L. F., Lian, T., & Eilifsen, C. (2011). Response-to-sample requirements in conditional discrimination procedures. *European Journal of Behavior Analysis, 12*, 505–522. Retrieved from http://www.ejoba.org.

Arntzen, E., Grondahl, T., & Eilifsen, C. (2010). The effects of different training structures in the establishment of condiitonal discriminations and subsequent performance on tests for stimulus equivalence. *The Psychological Record, 60*, 437–462.

Arntzen, E., & Holth, P. (1997). Probability of stimulus equivalence as a function of training design. *The Psychological Record, 47*, 309–320.

Arntzen, E., & Holth, P. (2000). Equivalence outcome in single subjects as a function of training structure. *The Psychological Record, 50*, 603–628.

Arntzen, E., Nartey, R. K., & Fields, L. (2014). Identity and Delay Functions of Meaningful Stimuli: Enhanced Equivalence Class Formation. *The Psychological Record, 64*, 349–360. doi: 10.1007/s40732-014-0066-3

Arntzen, E., Norbom, A., & Fields, L. (2015). Sorting: An alterternative measure of class formation. *The Psychological Record, 65*(2), 1–11. doi: 10.1007/s40732-015-0132-5

Barnes, D., & Keenan, M. (1993). A transfer of function through derived arbitary and nonarbitrary stimulus relations. *Journal of the Experimental Analysis of Behavior, 59*, 61–81. doi: 10.1901/jeab.1993.59-61

Buffington, D. M., Fields, L., & Adams, B. J. (1997). Enhancing equivalence class formation by pretraining of other equivalence classes. *The Psychological Record, 47*, 69–96.

Cowley, B. J., Green, G., & Braunling-McMorrow, D. (1992). Using stimulus equivalence procedures to teach name-face matching to adults with brain injuries. *Journal of Applied Behavior Analysis, 25*, 461–475. doi: 10.1901/jaba.1992.25-461

Dymond, S., & Rehfeldt, R. A. (2001). Supplemental measures of derived stimulus relations. *Experimental Analysis of Human Behavior Bulletin, 19*, 8–12.

Eilifsen, C., & Arntzen, E. (2009). On the role of trial types in tests for stimulus equivalence. *European Journal of Behavior Analysis, 10*, 187–202. Retrieved from http://www.ejoba.org.

Eilifsen, C., & Arntzen, E. (2011). Single-subject withdrawal designs in delayed matching-to-sample procedures. *European Journal of Behavior Analysis, 12*, 152–172. Retrieved from http://www.ejoba.org.

Fields, L., Arntzen, E., & Moksness, M. (2014). Stimulus Sorting: A quick and sensitive index of equivalence class formation. *The Psychological Record, 64*, 487–498.

Fields, L., Arntzen, E., Nartey, R. K., & Eilifsen, C. (2012). Effects of a meaningful, a discriminative, and a meaningless stimulus on equivalence class formation. *Journal of the Experimental Analysis of Behavior, 97*(2), 163–181.

Fields, L., Reeve, K. F., Rosen, D., Varelas, A., Adams, B. J., & Belanich, J. (1997). Using simultaneous protocol to study equivalence class formation: The facilitating effect of nodal number and size of previously established equivalence classes. *Journal of the Experimental Analysis of Behavior, 67*, 367–389. doi: 10.1901/jeab.1997.67-367

.

Fienup, D., & Dixon, M. (2006). Acquisition and maintenance of visual-visual and visual-olfactory equivalence classes. *European Journal of Behavior Analysis, 6*, 87–98. Retrieved from http://www.ejoba.org.

Grimm, L. J. (2011). *Emergence of equivalence relations: comparing sorting and match-to-sample procedures.* Retrieved from http://hdl.handle.net/2047/d20001017

Hove, O. (2003). Differential probability of equivalence class formation following a one-to-many versus many-to-one training structure. *The Psychological Record, 53*, 617–634. 152–172. Retrieved from http://opensiuc.lib.siu.edu/tpr/vol653/iss614/617/.

Keller, F. S., & Schoenfeld, W. N. (1950). *Principles of psychology: A systematic text in the science of behavior*. New York, NY: Appleton-Century-Crofts.

Lowe, C. F., Horne, P. J., Harris, F. D. A., & Randle, V. R. L. (2002). Naming and categorization in young children: vocal tact training. *Journal of the Experimental Analysis of Behavior, 78*, 527–549. doi: 10.1901/jeab.2005.31-04

Ludvigson, H. W., & Caul, W. F. (1964). Relative effect of overlearning on reversal and nonreversal shifts with two and four sorting categories. *Journal of Experimental Psychology, 68*(3), 301–306.

Mackay, H. A., Wilkinson, K. M., Farrell, C., & Serna, R. W. (2011). Evaluating merger and intersection of equivalence classes with one member in common. *Journal of the Experimental Analysis of Behavior, 96*(87–105). doi: 10.1901/jeab.2011.96-87

Michael, J. (1993). Establishing Operations. *The Behavior Analyst, 16*(2), 191–206.

Nartey, R. K., Arntzen, E., & Fields, L. (2014a). Two discriminative functions of meaningful stimuli that enhance equivalence class formation. *The Psychological Record, 64*, 777–789. doi: 10.1007/s40732- 014-0072-5

Nartey, R. K., Arntzen, E., & Fields, L. (2014b). Enhancement of equivalence class formation by pretraining discriminative functions. *Learning & Behavior, 43*, 20–31. doi: 10.3758/s13420-014-0158-6

Nedelcu, R. L., Fields, L., & Arntzen, E. (2015). Conditional discriminative functions of meaningful stimuli and enhanced equivalence class formation. *Journal of the Experimental Analysis of Behavior, 103*, 349–360. doi: 10.1002/jeab.141

Pierce, D. W., & Cheney, C. D. (2008). *Behavior Analysis and Learning* (4 ed.). New York, NY: Psychology Press.

Rosch, E., & Mervis, C. B. (1977). Children's sorting: A reinterpretation based on the nature of abstraction in natural categories. In R. C. Smart & M. S. Smart (Eds.), *Readings in child development and relationships* (2 ed., pp. 140–148). New York, NY: MacMillan.

Saunders, R. R., & Green, G. (1999). A discrimination analysis of training-structure effects on stimulus equivalence outcomes. *Journal of the Experimental Analysis of Behavior, 72*, 117–137.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth Cengage Learning.

Sidman, M. (1960). *Tactics of scientific research*. New York, NY: Basic Books, Inc.

Sidman, M. (1992). Equivalence relations: Some basic considerations. In S. C. Hayes & L. J. Hayes (Eds.), *Understanding verbal relations (pp. 15–27)*. Reno, NV: Context Press.

Sidman, M. (1994). *Equivalence relations and behavior: A research story*. Boston, MA: Authors Cooperative.

Sidman, M., & Tailby, W. (1982a). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37*, 5–22.

Sigurdardottir, Z. G., Mackay, H. A., & Green, G. (2012). Stimulus equivalence, generalization, and contextual stimulus control in verbal classes. *The Analysis of Verbal Behavior, 28*, 3–29. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3363409/.

Watt, A., Keenan, M., Barnes, D., & Cairns, E. (1991). Social categorization and stimulus equivalence. *The Psychological Record, 41*, 33–50. http://psycnet.apa.org/psycinfo/1991-17495-17001.

Table 1

Groups and conditions

Assigned	Phases				
Group 1	Pre-SRT	TBR	SRT-1	MTS (1a+1b)	SRT-2
Group 2	Pre-SRT	TBR	MTS (1a+1b)	SRT-1	MTS (2a+2b)

Note. The table shows which conditions the two groups will be exposed to. Pre-SRT = The pre class-formation sorting test; TBR = Training of baseline relations; SRT-1 = The first post class formation sorting test; SRT-2 = The second post class-formation sorting test; MTS = The Matching-To-Sample test; 1a = The first block of the first MTS test; 1b = The second block of the first MTS test; 2a = The first block of the second MTS test; 2b = The second block of the second MTS test.

Table 2

Summarized results from the experiment

	Gender	Age	PN	Pre-SRT						TBR	SRT-1					MTS (1a+1b)					SRT-2					
Group 1	F	21	15101	221	123	221				720	**500**	**050**	**005**			**100**	**100**				**500**	**050**	**005**			
	F	21	15119	221	221	113				660	**500**	**050**	**005**			**99**	**100**				**500**	**050**	**005**			
	F	23	15103	331	213	011				960	**500**	**050**	**005**			**96**	**100**				**500**	**050**	**005**			
	F	20	15113	221	222	112				1020	**500**	**050**	**005**			65	71				**500**	**050**	**005**			
	F	24	15117	131	212	212				1200	**500**	**050**	**005**			53	54				122	212	221			
	F	20	15105	112	112	310	021			1200	**500**	020	**005**	030		79	73				400	030	**005**	120		
	F	23	15107	221	221	102	011			540	300	**050**	004	201		88	87				300	**050**	205			
	F	22	15111	303	121	010	121			720	300	**050**	003	200	002	53	52				400	040	103	012		
	M	33	15109	211	330	003	011			1440	112	112	220	111		47	33				331	021	203			
	M	22	15115	020	110	011	200			960	221	211	121	002		43	38				112	020	200	110	002	111

	Gender	Age	PN	Pre-SRT						TBR	MTS (1a+1b)			SRT-1					MTS (2a+2b)		
Group 2	F	21	15118	101	200	011	011	020	212	540	**98**	**100**		**500**	**050**	**005**			**100**	**100**	
	F	26	15122	033	411	111				660	**97**	**99**		**500**	**050**	**005**			**100**	**100**	
	F	40	15102	320	122	113				540	**95**	**98**		**500**	**050**	**005**			**98**	**100**	
	F	20	15104	122	122	311				1080	65	56		**500**	**050**	**005**			56	54	
	F	25	15120	121	321	113				1140	61	58		**500**	**050**	**005**			65	58	
	F	23	15121	111	111	120	102	111		600	46	62		**500**	**050**	**005**			**97**	**99**	
	F	26	15106	130	112	211	102			1740	87	83		**500**	030	003	020	002	88	83	
	F	27	15108	301	012	021	221			1080	71	87		400	150	**005**			73	73	
	F	26	15110	122	212	221				1380	58	59		400	031	003	121		57	61	
	F	44	15114	021	201	111	012	210		2100	54	46		021	201	112	121	100	42	40	

Note. The table shows the data from each participant regarding all phases of the experiment. PN = Participant number; MTS = The Matching-To-Sample test; 1a = The first block of the first MTS test; 1b = The second block of the first MTS test; 2a = The first block of the second MTS test; 2b = The second block of the second MTS test; SRT = The post class-formation sorting test; Numbers beneath the MTS tests = Percentage correct performance (i.e., all directly trained, symmetrical, transitive, and equivalence relations per participant are added together and then divided on the four types of relations which yields an overall stimulus equivalence score), whereas bolded numbers indicate percentage correct performance and in accordance with stimulus equivalence. Numbers regarding SRTs = The number of stimuli from each class that is laid in each participant-defined cluster (e.g., Participant 15102s performance 320 means that three of the Class-1 stimuli, two of the Class-2 stimuli, and none of the Class-3 stimuli was clustered together); Bold font = Highlighted the performances where the participant-defined classes matched the experimenter-defined classes (e.g., the performance in Participant 15118 which yielded 500, 050, 005, was an example where the participant sorted five Class-1 stimuli in one cluster, five Class-2 stimuli in another cluster, and five Class-3 stimuli in a third cluster, from which all clusters matched the experimenter-defined classes).

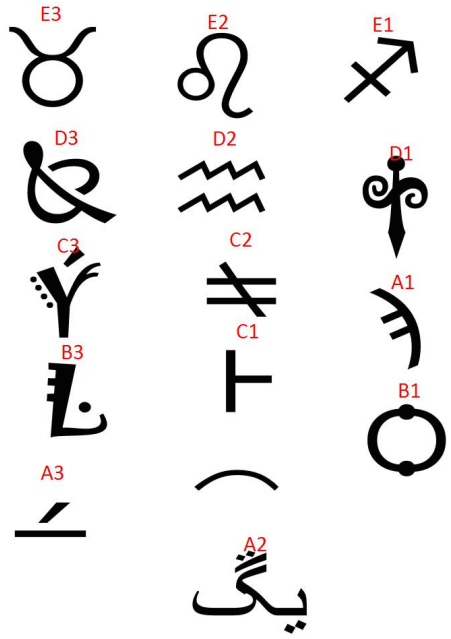

Figure 1. The figure shows the performance on the post class-formation sorting test by participant 15108. All the stimuli from Class 3 are clustered together in a way that matched the experimenter-defined class. However, Class 1 are not complete with one stimuli missing, in which was sorted in Class 2 by the participant. In Table 1, this can be seen as a 400 150 005 performance. The red letters and numbers above each stimulus were not visible to the participants during the experiment.

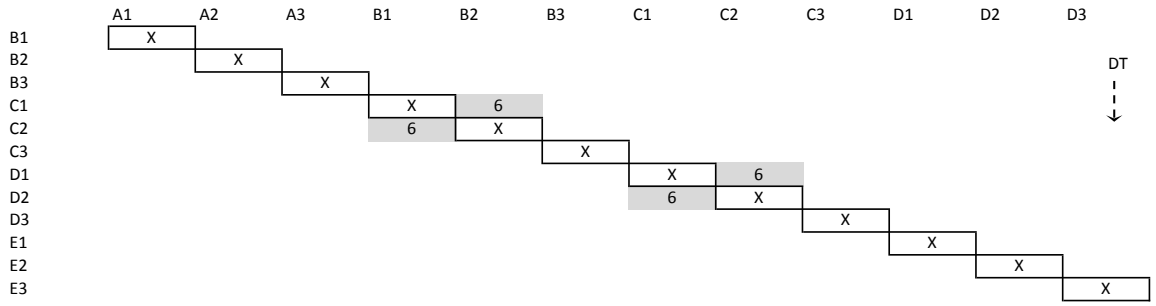

Figure 2. The figure shows the error rate for participant 15108, who was the only participant by which showed participant-defined classes that persisted throughout two subsequent test blocks. This particular illustration shows the testing of directly trained trials on the two last test blocks (MTS 2a + 2b) and must be read from the top-to-bottom. The participant responded in accordance to participant-defined classes in which was systematically repeated across the two last test blocks in particular. The grey marked cells shows where the participant was responded incorrect (i.e., on each occasion when the C1-stimuli was involved, but also when the C2-stimuli was involved).

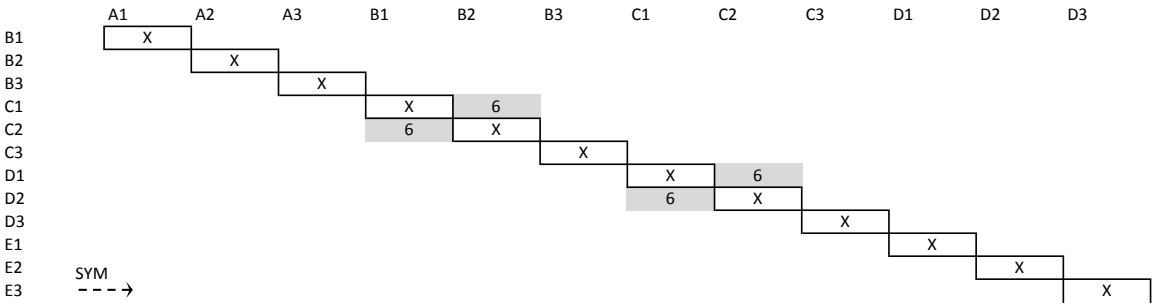

Figure 3. The figure shows the error rate for participant 15108 during the second MTS test. This particular illustration shows the symmetry trials on the two last test blocks (MTS 2a + 2b) and must be read from the left-to-right axis. The participant responded in accordance to participant-defined classes in which was systematically repeated across the two last test blocks in particular. The grey marked cells shows where the participant was responded incorrect (i.e., on each occasion when the C1-stimuli was involved, but also when the C2-stimuli was involved).

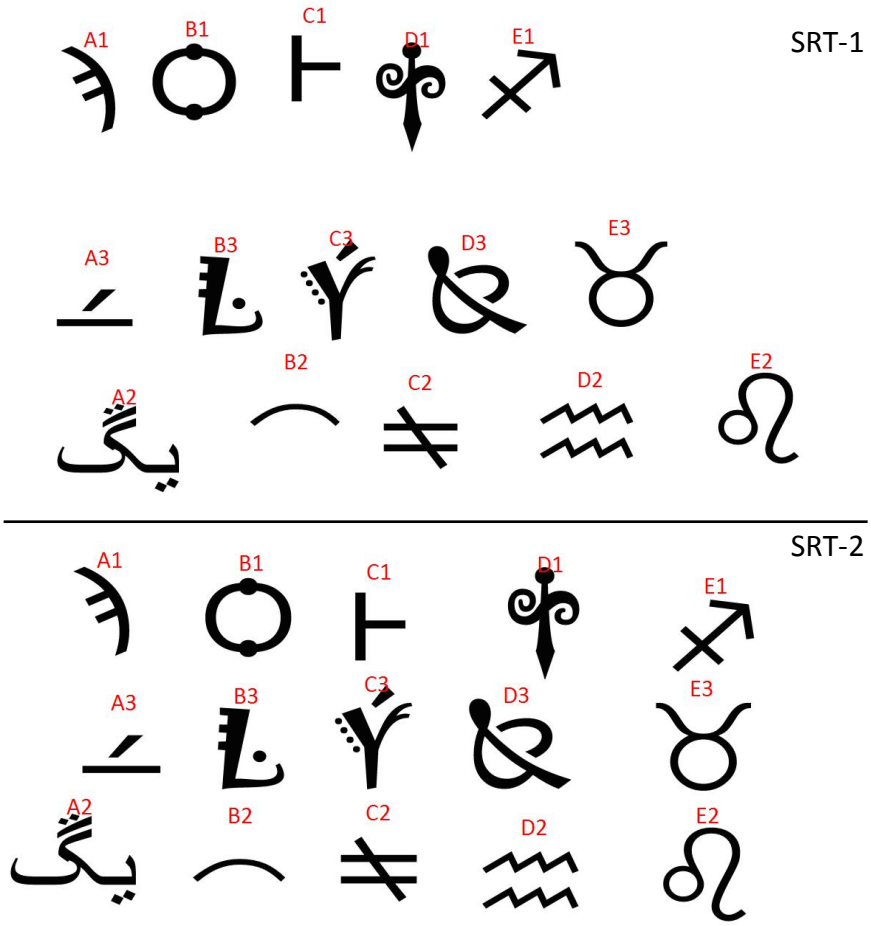

Figure 4. The figure shows the performance on SRT-1 and SRT-2 for Participant 15113. This participant did not respond in accordance with stimulus equivalence in the MTS test administered between the two sorting tests. Note that the stimuli sorted are in exactly the same order as the relations that were established in training (i.e., the baseline relations) in both sorting tests. The red letters and numbers above each stimulus were not visible to the participants during the experiment. The red letters and numbers above each stimulus were not visible to the participants during the experiment.

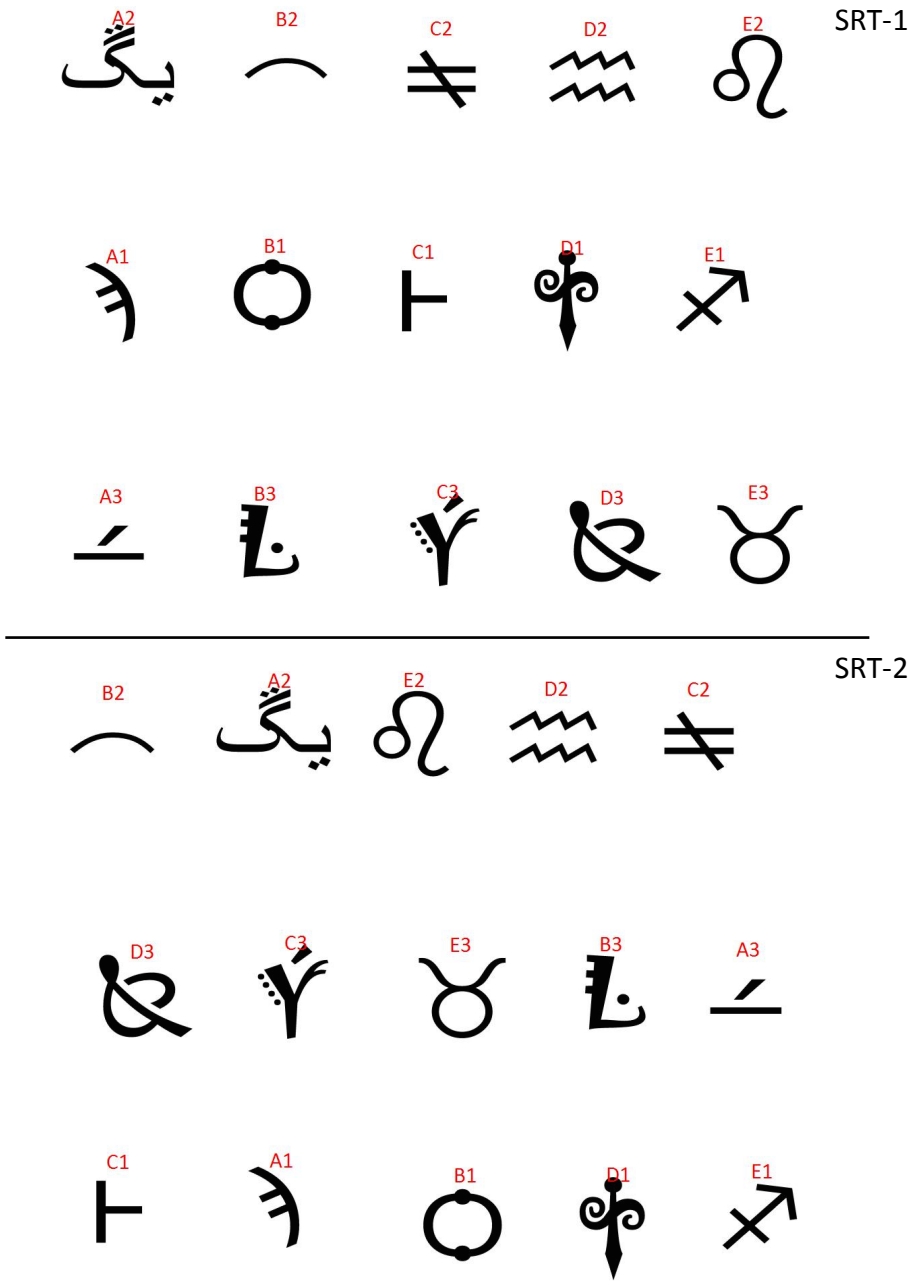

Figure 5. The figure shows the performance of Participant 15119 on SRT-1 (i.e., upper panel) and SRT-2 (lower panel). As can be seen, SRT-1 shows sorting that are exactly as the relations was trained. In SRT-2, the participant shows a more random order of the relations, however, all of which belonging to separate classes that matched the experimenter-defined classes. Note that this participant did show SE responding in the MTS test which was administered between these two sorting tests. The red letters and numbers above each stimulus were not visible to the participants during the experiment.

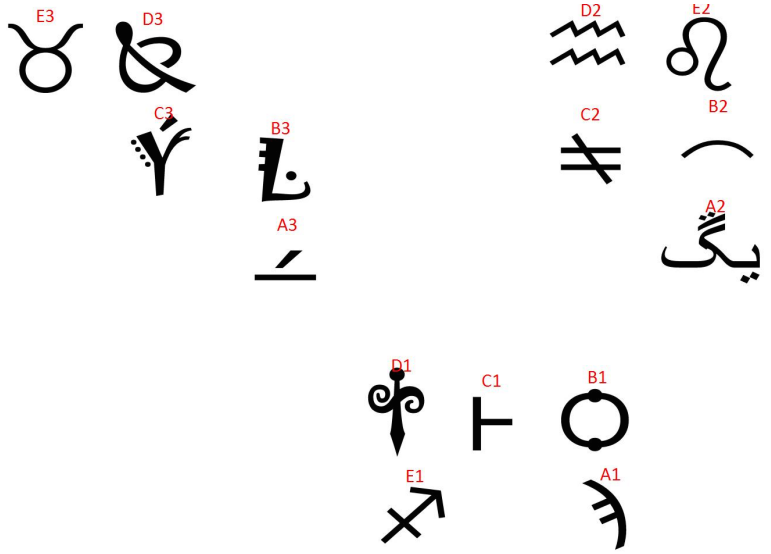

Figure 6. The figure shows the performance on SRT-1 in participant 15122. This participant received only one sorting test, and it was administered between two MTS tests. Both MTS tests were responded to in accordance with stimulus equivalence. As can be seen from the figure, the stimuli are sorted in much of the same manner as seen in Figure 2 and the upper panel of Figure 3, where the stimuli are sorted as they are trained (e.g., A3-B3-C3-D3-E3, not D3-A3-E3-B3-C1). The red letters and numbers above each stimulus were not visible to the participants during the experiment.

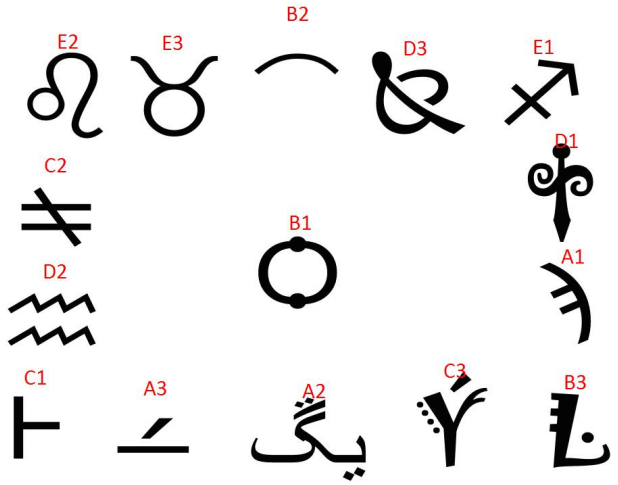

Figure 7. The figure illustrates a case where the participant performed in a way could be interpreted in many different ways, but in which could not be regarded as correct in either of the ways possible to interpret. In such cases, the Interobserver Agreement was regarded as irrelevant due to the lack of correct performance anyhow. The red letters and numbers above each stimulus were not visible to the participants during the experiment.