The Psychological Record
https://doi.org/10.1007/s40732-020-00377-3

ORIGINAL ARTICLE

Yield as an Essential Measure of Equivalence Class Formation, Other Measures, and New Determinants

Lanny Fields[1] · Erik Arntzen[2] · Erica Doran[3]

Abstract
"Yield," the percentage of participants in a group who form a set of equivalence classes, has been used very broadly to identify the effect of different training protocols on class formation and expansion, identify variables that enhance the immediate emergence of these classes, and characterize the differential relatedness of class members. In addition, yield is now being used to document the formation of educationally relevant equivalence classes. To further understand the value of using yield, we considered six possible criticisms of its use to study equivalence classes. Upon analysis, each criticism was supported; instead, each disclosed a nonyield factor that could play a critical role in the measurement of class formation but has not yet been explored experimentally. Finally, yield cannot be replaced with trial-based measures of responding or vice versa; rather, both types of measures are needed to obtain a comprehensive understanding of equivalence class formation.

Keywords equivalence class formation · yield · stimulus control topographies · immediate emergence · delayed emergence · postclass formation stimulus relatedness

Overview

An equivalence class is a set of perceptually disparate stimuli that have come to function interchangeably. After training some of the stimulus–stimulus relations in a potential class, the class has been formed when all of the untrained relations occasion the mutual selection of each other. When these performances emerge with the first presentation of the untrained relational probes, the class has emerged on immediate basis. In such an instance, accuracy of responding shifts in one step from essentially random responding before training to class-indicative responding immediately after training, i.e., there are no intermediate levels of test responding that occur that would permit the tracking of a gradual transition in performance. Thus, the only way of tracking the effect of a particular training package on the immediate emergence of class formation is by computing the percentage of participants in a group who form the classes, which is called "yield." Further, the effect of an independent variable on class formation can be characterized by demonstrating how the values of yield are influenced by particular values of the independent variable.

Yield is now widely used to document the effects of many variables on equivalence class formation. Thus, it would appear that the use of yield has become settled practice. We are of the opinion, however, that its use needs further elucidation. This article will consider six possible criticisms of the use of yield to study equivalence class formation. Each will be considered to determine whether it supports the view that yield should not be used measure likelihood of equivalence class formation. In addition, we will consider the unexpected implications of each criticism for the study of as yet unexplored non-yield-based factors that might influence class formation. Finally, we will argue that a comprehensive understanding of equivalence classes can be provided only by a consideration of yield, the gradual emergence of equivalence classes, and of the differential relatedness of stimuli in fully formed classes.

✉ Lanny Fields
Lanny.fields1@gmail.com

[1] Queens College and The Graduate Center, The City University of New York, New York City, USA

[2] Oslo Metropolitan Universit, Oslo, Norway

[3] Queens College and Queensboro Community College, The City University of New York, and St. John's University, New York, USA

Published online: 31 March 2020

Equivalence Classes, Illustrated, Defined, and Measured

Assume that a person does not know of the many representations of "FIVEness" such as: *five, cinco, 5, xxxxx* (i.e., five items), and *0101* (i.e., the binary representation of 5). For her to demonstrate an understanding of FIVEness, she would have to recognize that all of the representations of FIVE are discriminable from all of the representations of other numbers such as THREEs (*three, tres, 3, xxx,* and *0011*), SEVENs (*seven, siete, 7, xxxxxxx,* and *0111*), and *EIGHTs* (*eight, ocho, 8, xxxxxxx,* and *1000*), etc. In addition, all of the representations of FIVE would have to be recognized as being related to and interchangeable with each other. Both of these goals can be achieved by training a small number of baseline relations among the representations of FIVE such as five–cinco, cinco-5, 5–xxxxx, and xxxxx-0101. These linked baseline relations should logically produce many untrained or relations that include cinco-five, 5-cinco, xxxxx-5 and 0101-xxxxx, which are 0-node symmetrical relations probes; five–5, cinco-xxxxx and 5-0101, which are 1-node transitive relations probes; 5-five, xxxxx-cinco and 0101-5, which 1-node equivalence relations probes; five-xxxxx and cinco-0101, which are 2-node transitive relations probes; xxxxx-five and 0101-cinco, which are 2-node equivalence relations probes; five-0101, which is a 3-node transitive relation probe; and 0101-five, which is a three-node equivalence relation probe.

If each of these probes resulted in the selection of the second stimulus in a pair when presented with the first, all of the performances would document the properties of symmetry, transitivity, and equivalence (i.e., the combined effects of symmetry and transitivity) all of which define equivalence (Sidman & Tailby, 1982). Thus, all of the stimuli in the set would be acting as members of an equivalence class. In addition, if a participant was then trained to say "/THEENKOH/" when presented with 5, most likely, she would always say "/THEENKOH/" when presented with any of the other representations of FIVE, and would not say it when presented with any of the representations of THREEs, SEVENs, or EIGHTs, etc. Thus, the equivalence class would also be functioning as a response transfer network. A more detailed characterization of the emergent or derived relations in an equivalence class can be found in Fields and Verhave (1987).

From a more formal perspective, an equivalence class like that presented in the preceding paragraph, contains N perceptually disparate stimuli (five in the example above) and N^2 relations among those stimuli (or 25 in the example above). A class can be formed by the training of (N-1) relations between the stimuli, all of which are called baseline relations (Fields & Verhave, 1987). For example, in a set of five stimuli referred to as A, B, C, D, and E, one set of baseline relations would be AB, BC, CD, and DE. Such a set of linked relations can give rise to the emergence of all the remaining (N^2-N+1) ordered pairs or probes in the set. If all or most of these probes are presented in a test block and all of them produce class-indicative responding, these emergent performances document the formation of an equivalence class that consists of the A, B, C, D, and E stimuli.

If this occurs in the first or second administration of the test block, it documents the "immediate" emergence of the equivalence class (e.g., Saunders, Chaney, & Marquis, 2005). The immediate emergence of an equivalence class occurs on an "all-or-none basis." As a measure, immediate emergence is a static phenomenon that does not change across test blocks. When viewed in the context of a group of participants, immediate emergence can be indexed by the percentage of participants who have formed the classes in the first or second test block, and has been referred to as yield (Fields et al., 2000). Thus, yield has been used to identify variables that influence the immediate emergence of equivalence classes for groups of participants, some examples of which are included in the following section.

Yield and Variables that Influence Equivalence Class Formation

Choices per trial, training protocol, class size, and nodal number Saunders et al. (2005) studied how class formation by senior citizens was influenced by the number of comparisons, delay durations, training structure, and class size. Three- and four-member classes were formed using Sample as Node (also called One to Many), Comparison as Node (also called Many to One), or Linear Series training structures. For each condition, the trials contained either two, three, or four comparisons as choices. In Experiment 1, trials were conducted using trials in which the sample stimuli remained on after the presentation of the comparison stimuli: the no-delay condition. Experiment 2 replicated Experiment 1 with one exception: the sample stimulus was terminated at the same time as the comparison sets were presented—the 0-s delay condition. In the no-delay condition, all training structures resulted in similar intermediate yields, with little influence of number of choices per trial or class size. In the 0-s delay condition, a different pattern of yield occurred across training structures. 100% yields were obtained when for Sample as Node or Comparison as Node training structures, for either class size. In contrast, much lower yields were observed when the classes had Linear Series training structures, and yields were lowest when using four comparisons per trial. Thus, yield documented the effects of four of the parameters used to form the equivalence classes.

Adams, Fields, and Verhave (1993) studied the effects of two training and testing protocols on the formation of one-node, three-member classes, and then their expansion to two-node, four-member, and finally to three-node five-member classes. In one group. all training and testing were conducted using the simple-to-complex protocol (STC), and for the other, all training and testing was conducted using the complex-to-simple protocol (CTS). In the STC protocol, each baseline relation is trained separately, and each type of derived relation is introduced serially, and is interleaved with the training of the baseline relations. In addition, the derived relations probes are introduced in increasing order of complexity: symmetrical relations probes first, transitive relations probes next, and equivalence relations probes last. Finally, for the transitive and equivalence relation probes, they are introduced from the smallest to the largest nodal separations, respectively. In contrast, the CTS protocol involves the serial training of the baseline relations, after which participants are presented with the equivalence probes that contain the maximal nodal separation first, followed by the presentation of less complex probes.

All participants in the STC group formed the three-member classes, and then expanded the class sizes to four- and five-member classes. In all cases, 100% yields were obtained for the formation of the classes in all sizes. In the CTS group, all participants formed the three-member classes (100% yields), fewer formed the four-member classes (lower yields), and even fewer of the participants who formed the four-member classes showed expansion of class size to five-members (lowest yield). Thus, yield quantified the interactive effects of STC and CTS protocols on the likelihood of class formation and class expansion.

Equivalence classes can also be studied in the context of a third training and testing routine called the Simultaneous protocol (SIM; Buffington, Fields, & Adams, 1997). In the SIM protocol, all baseline relations are formed concurrently after which all of the derived relations probes are presented in a given test block, i.e., concurrently. In general, yields are much lower when training and testing are conducted under the SIM protocol than the STC protocol. For example, Fienup, Wright, and Fields (2015) found that the immediate emergence of three- and four-member classes was more likely to occur during the STC protocol than during the SIM protocol. In addition, Fields et al. (1997) found that the percentage of participants who formed new three-node, five-member equivalence classes (yield) was influenced by the previously formed equivalence classes. In particular, the likelihood of forming new three-node, five-member equivalence classes under the SIM protocol was a direct linear function of the size (three through seven members) and number of nodes (one through five) in other equivalence classes that had been previously formed under the STC protocol

Enhancement of equivalence class formation The immediate emergence of equivalence classes was enhanced by variables that were inherent components of prospective classes such as (1) their nodal structures (Arntzen & Holth, 2000; Fields, Hobbie-Reeve, Adams, & Reeve, 1999); (2) the inclusion of a meaningful stimulus as a class member (Fields & Arntzen, 2018; Fields, Arntzen, Nartey, & Eilifsen, 2012); (3) the location of a meaningful stimulus in the structure of the class (Nartey, Arntzen, & Fields, 2015b); (4) the order of introducing a meaningful stimulus during the training of the baseline relations (Nartey et al., 2015b); and (5) the number of meaningful stimuli in a to-be-formed class (Mensah & Arntzen, 2017).

In addition, the immediate emergence of an equivalence class was enhanced by any one of following stimulus control functions that were acquired by one meaningless stimulus before its inclusion in a to be formed equivalence class: These included (1) simultaneous and successive simple discriminative functions, alone and in combination (Fields, Arntzen et al., 2012; Nartey, Arntzen, & Fields, 2015a); (2) identity and arbitrary conditional discriminative functions, using either simultaneous or delayed matching (Arntzen, Nartey, & Fields, 2014; Arntzen, Nartey, & Fields, 2015a); (3) the overtraining of the simple successive discriminative function (Travis, Fields, & Arntzen, 2014); and (4) the number of arbitrary conditional relations that share a stimulus that is to become a member of the target equivalence class (Nedelcu, Fields, & Arntzen, 2015).

Finally, some quantitative values of these variables produced yields that were essentially the same as those produced when the classes included only one meaningful stimulus. Thus, the acquired stimulus control functions could account for the class-enhancing effects of a meaningful stimulus that is included as a member of an equivalence class. To sum up, yield has played a significant role in the discovery of independent variables that enhance the immediate emergence of equivalence classes (Arntzen, Nartey et al. 2015a).

Relatedness of stimuli in equivalence classes According to Sidman (1994, 2000), once an equivalence class has formed, all of the stimuli are interchangeable and by implication equally related to each other. In contrast, Fields and Verhave (1987) proposed that the stimuli in an equivalence class could concurrently be differentially related to each other based on their nodal separation. This view has been supported by a growing number of experiments that have shown that the stimuli in an already formed equivalence class are interchangeable when assessed with cross-class tests, and are differentially related to each other when presented a variety of other types of tests. Some contain stimuli all of which come from the same class (within-class preference tests), others pit responses trained to different members of the same class against each other (dual option response-transfer tests), and others that used traditional MTS trials but measure response speed produced by different types of trials, and the same types of trials that vary in terms of nodal distance (response-speed tests).

Within-class preference tests A within-class preference test trial contains a sample from one class with two comparisons that differ from the sample by a different number of nodes. For example, after forming an ABCDEFG class, responding on the within class preference tests have shown that the relatedness of class members to be an inverse function of the nodal separation of stimuli holding relational type constant (Moss-Lourenco & Fields, 2011; Wang, Dack, McHugh, & Whelan, 2011); that transitive relations are preferred to equivalence relations, holding nodal number constant (Doran & Fields, 2012); and that preference is a joint function of relational type and nodal separation Albright, Fields, Reeve, Reeve, & Kisamore, 2019). In each of these studies, because the same pattern of responding was produced across participants, that pattern defines a constant performance. It therefore follows that the prevalence of the pattern can be summarized as a statement of yield such as ". . . the same pattern of responding occurred across 95% of the tests and by 100% of the participants."

Dual-option response-transfer tests Fields, Landon-Jimenez, Buffington, and Adams (1995) formed three-node, five-member equivalence classes with structures represented as A➔B➔C➔D➔E. Thereafter, participants were trained to make different responses in the presence of the A and E stimuli from the same class. In a final transfer test, each of the class members was presented alone, and the likelihood of making the A- or E-response was measured for each class member. Generalization of these responses was an inverse function of the number of nodes that separated the B, C, and D stimuli from the A and E stimuli.

Fields and Watanabe-Rose (2008) formed four-node, six-member equivalence classes with structures represented as A➔B➔C➔D➔E➔F after which different responses were trained to the C and D stimuli. In the following transfer test, the C-response transferred completely to the B and A stimuli whereas the C-response transferred completely to the E and F stimuli, thereby bifurcating the six-member class into two three-member classes with class membership dictated by nodal structure ABC and DEF. In this experiment, the intactness of the initial class was confirmed at the completion of the transfer test, which showed that the bifurcation did not interfere with the intactness of the original six-member class.

The two studies mentioned in this section used dual option within class tests to assess the relatedness of class members after the classes had emerged. The results of both provide further support for the view that the stimuli in an equivalence class are differentially related to each other when assessed on a within class basis and are interchangeable when assessed on a cross class basis. Further confirmation of these outcomes has also been provided by the outcomes of semantic differential tests conducted with stimuli in multi-nodal equivalence classes (Bortoloti & de Rose, 2009).

Response-speed tests Spencer and Chase (1996) formed five-node, seven-member classes with a class structure represented by A➔B➔C➔D➔E➔F➔G. Once formed, response speed (the reciprocal of reaction time) was measured for each relational type and or each nodal distance for the transitive relations and the equivalence relations. Response speed was fastest for the baseline relations, and was in inverse function of the number of nodes that separated the stimuli in the derived relations. Thus, response speed showed differential relatedness among the stimuli in fully formed equivalence classes.

Fields et al. (1995) also measured reaction times produced by the A through E stimuli in the dual option transfer tests and found that they were fastest in the presence of the A and E stimuli, slower in the presence of the B and D stimuli, and longest in the presence of the C stimuli. These chronometric data were also an inverse function of nodal distance in the dual option test.

Although these chronometric measures documented postclass-formation effects of nodal distance, these outcomes have not always been obtained. For example, Tomanari, Sidman, Rubio, and Dube (2006) added a contingency to maximize short latencies while training the baseline relations for the equivalence classes. After class formation, similar latencies were produced by the sample stimuli in the symmetry and equivalence probes. In contrast, the comparison stimuli produced slightly longer latencies to the equivalence probes than to the symmetry probes for some participants, but not for the others. Thus, special contingencies of reinforcement influenced the ability of response latency to reflect the effects of a structural parameter of an equivalence class.

Finally, as with yield, it could also be said that reaction time or response speed does not measure behavior, but rather is a proxy for behaviors that involve the processing of the information contained in the sample and comparison stimuli that define an MTS trial. Perhaps another line of research that would further clarify the process of forming and equivalence class or that would characterize the relatedness of stimuli in an already formed class would involve the measurement of actual behaviors that occur during the temporal epoch that is defined the duration of a reaction time.

Section summary

In each of the experiments mentioned in this section, the same complex set of performances were observed for all participants. Thus, these outcomes could be summarized by yield statements that would confirm the reliability of the findings across tests and participants. It also has to be emphasized that these measures are not replacements for yield. Indeed, it is only after the classes have been formed that measures can be meaningfully acquired to characterize the differential relatedness of the stimuli in an equivalence class. Yield on the one hand, and within class measures on the other, are complementary rather than interchangeable with or substitutable for each other.

Validity and Utility of Yield

All these findings demonstrate that yield has played a substantial role in the identification of variables that have influenced equivalence class formation and the relatedness of stimuli in an equivalence class. The use of yield can also be viewed from two broader perspectives. First, yield has been used routinely in double blind studies to (1) compare the efficacy of a new drug relative to a placebo treatment, and (2) evaluate the relative efficacies of different drug treatments for particular medical ailments.

Second, a primary goal of an educational entity is to use pedagogical protocols that result in the rapid acquisition of a target skill by most of the students in a class. In the last 10 years, an increasing number of experiments have used yield (1) to document the efficacy of *equivalence based instruction* (EBI) for the establishment of classes of academically relevant materials with college-level students (Fields et al., 2009; Fienup et al., 2015; Spear & Fields, 2015; Walker & Rehfeldt, 2012), and (2) for children with severe intellectual impairments (Arntzen, Halstadtro, Bjerke, & Halstadtro, 2010; Arntzen, Halstadtro, Bjerke, Wittner, & Kristiansen, 2014a; De Souza & Rehfeldt, 2013).

In both of these contexts, yield has been used to document the efficacy of medical or educational protocols. Thus, the use of yield to document the immediate emergence of equivalence classes would appear to be in well-regarded scientific company.

Six Critiques of Yield

Although yields are now widely used to explore equivalence classes, there are still issues that can be raised regarding its validity. In particular, it could be argued that yield (1) does not "measure" behavior; (2) can lead to inappropriate cross-experiment comparisons; (3) does not take "near misses" into consideration; (4) does not prompt use of sensitive statistical measures and is not sensitive to sample size; (5) does not identify variables that influenced individual performances; and (6) does not identify sources of stimulus control that influence class formation. Before considering these critiques, we will examine the many ways in which class formation has been defined for individual participants.

Quantitative definitions of equivalence class formation In general, yield is defined as the percentage of participants in a group who formed equivalence classes. Yield is necessarily determined by the way in which class formation is measured (for an early systematic analysis, see Doran, 2009). Thus, yield will be influenced by the way class formation is defined for individual subjects. The most frequently used measure of class formation has been the percentage of trials in a test block that produce class indicative comparison selections (e.g., Arntzen, 2012; Goyos, 2000; Sidman, 1971; Sidman & Cresson, 1973). Within that context, class-formation criterion has been set at accuracies that ranged from 100% to at least 75% of trials in a block that produced class indicative responding. Most studies, however, have used either 100%, at least 90%, or at least 80% accuracy.

One recent experiment defined class-formation criterion by the administration of only two test blocks, each of which contained 24 trials. Class formation was documented by the selection of correct comparisons on at least 22 of the 24 trials in each block—at least 91.7% correct (Bortoloti, Rodrigues, Cortez, Pimentel, & de Rose, 2013). Yet another experiment set an overall class-formation criterion of at least 90% for the entire block and included at least 80% accuracy for each of the derived relations probes in the block (e.g., Steele & Hayes, 1991). For yet other experiments, the class-formation criterion was defined by a combination of percentage correct and number of consecutive blocks that produced mastery. For example, Dougher (1994) defined it as at least 95% correct on six consecutive blocks, whereas Saunders, Saunders, Williams, and Spradlin (1993) defined it as being 100% correct in a given block or at least 90% correct in two consecutive blocks. In other studies, the class-formation criterion also required the occurrence of no more than one error on any derived relation probe (Spencer & Chase, 1996; Taylor & O'Reilly, 2000; Vie & Arntzen, 2019). Class formation has also been defined in terms of the number of consecutively administered trials that occasioned class-indicative selections (Devaney, Hayes, & Nelson, 1986; Fienup et al., 2015; Wulfert, Dougher, & Greenway, 1991).

With other experiments, the criteria for defining class formation used different mixes of relational types. For example, Eikeseth and Smith (1992) used all trial types: baseline, symmetry, transitivity, and equivalence. Cullinan, Barnes, and Smeets (1998) used all derived relations probes but not baseline relations. Finally, Bortoloti et al. (2013) formed three 3-node, five-member classes by training AB, AC, CD, and DE, which produced a training structure represented as B←A→C→D→E. Class formation was assessed with the administration of only two probe types, the most extended three-node equivalence relations that could emerge in each such class, BE and EB. In that study, each probe was presented in one block that contained 24 trials, 8 per class. Class formation was defined by the occurrence of at least 22 correct trials in each of the BE and EB blocks.

Finally, some experiments did not specify an explicit criterion for class formation. Rather, class formation was defined as performances that were "at or near 100%" (Lazar, 1977) or "near errorless" (Slotnick, & Silberburg, 1993). This information provides a context for considering the six critiques.

1. Yield does not "measure" behavior

It could be argued that yield is flawed because it does not measure behavior. The validity of that critique depends on the quantitative value of the mastery criterion used to define class formation. We will consider this notion under two conditions: 100% or near 100% accuracy and lower levels of accuracy.

Near 100% accuracy as mastery criteria If class formation is defined by 100% accuracy, all the relational probes, whether previously trained or emergent, produce the same level of class-indicative responding. In this case, yield reflects percentage of participants who perform in precisely the same manner in the presence of all probes. As such, yield is a valid proxy for perfect class-indicative performances across trial types and participants. In this case, yield is a valid measure of a participant's behavior. A similar argument can be made for mastery criteria that approximate 100% accuracy (e.g., at least 95% accuracy). As with 100% mastery, yield based on accuracies that approximate 100% can be used as a valid measure of behavior.

Lower accuracy as mastery criteria If the class-formation criterion is set at a much lower accuracy level, the class-consistency of responding across trial types and classes can become quite variable, as noted by Sidman (1987) and more recently by Arntzen (2012). For example, assume that class formation is defined by the evocation of class-consistent responding by the baseline, symmetry, transitivity, and equivalence probes on at least 80% of the trials in each of three classes. For one participant, the test trials for each class could produce 80% accuracy. For another participant, the probes for potential classes 1, 2, and 3 could produce 100%, 100%, and 40% accuracies. For each participant, averaging across all classes would produce the same overall accuracy of 80%. Thus, the performances of both participants would be included in the determination of yield.

Such an aggregation would be inappropriate because each average would reflect a different performance profile across classes. Thus, the inclusion of both outcomes would not provide a meaningful measure of yield. This problem, however, is not about the use of yield to stipulate class formation. Rather, it is about the setting of the mastery criterion used to define class formation at an insufficiently low level. Although this argument has been presented in the context defining class formation for individual participants (Sidman, 1987; Arntzen, 2012), it was not linked to the use of yield, as just considered.

Identifying an optimal mastery level for defining yield What then is the lowest class-formation criterion that should be used to define equivalence class formation? This question has not yet been addressed empirically. One potential solution to the problem would be to identify the lowest class-formation criterion that would produce an equivalence class that had the same functional properties of a class that was defined by a class-formation criterion of 100% accuracy. For example, that might be discovered by conducting a response transfer test after the establishment of classes with groups that used different class-formation criteria such as 100%, at least 95%, 90%, 85%, and 80% accuracy. One outcome might be that the class-formation criteria of 100% and 95% resulted in 100% response transfer, whereas the remaining criteria (at least 90%, 85%, and 80%) resulted in 78%, 23%, and 2% transfer. Because classes defined by a criterion of at least 95% accuracy resulted in complete response transfer, these classes would have the same functional properties as classes defined by 100% mastery. In this example, it would be appropriate to use at least 95% accuracy in a test block as the criterion for defining equivalence class formation.

In addition, the outcomes of a such a study would imply that the class-formation criteria that do not result in maximal transfer of function should not be used to define an equivalence class. This can be exemplified by a recent experiment that used a class-formation criterion of at least 80% accuracy was used to define equivalence classes that contained the names, definitions, and examples of three types of logical errors (Ong, Normand, & Schenk, 2018). Thereafter, little transfer occurred to another test that was designed to assess an understanding of the same types of logical errors. In that experiment, it would be interesting to determine whether substantial transfer to the follow-up test would have occurred by using a more stringent class-formation criterion such as "at least 95% accuracy."

Should equivalence be defined by function transfer? One final comment. This example shows how a response transfer outcome can be used to identify the minimal MTS based accuracy level that should be used to define the emergence of an equivalence class. This does not, however, imply that response transfer can be used to define an equivalence class because the performances produced by a transfer test cannot document the properties of reflexivity, symmetry, transitivity, or the combination of the latter two properties. Indeed, the proposed experiment mentioned above would probably clarify the conditions under which performances produced by MTS based tests of derived relations and by response transfer tests are consistent with each other (but see Wirth & Chase, 2002).

2. Yield can lead to inappropriate cross-experiment comparisons

It could be argued that yield is flawed because it can lead to inappropriate comparisons of the outcomes across experiments that have used different criteria to define class formation. This point was addressed in the prior section while considering the quantitative values used to define mastery, assuming all other parameters were constant. It will be considered from two other perspectives in this section.

As noted above, equivalence classes have been defined by the responses produced by different types and mixes of relational probes. Thus, different yields could be attributed to (1) differences in the values of an independent variable that distinguishes experiments, (2) different mixes of probes that were used to define class formation, and/or (3) the number or proportion of derived relations probes used to compute mastery.

These interpretative options imply that the criticism of yield is misdirected. Rather, they should be focused on determining how class formation is influenced by the types and number of relational probes used to define mastery. To date, most experiments have documented class formation with the presentation of a full battery of (N^2-N+1) emergent relations probes that assess symmetry, transitivity when available, and equivalence, where N is equal to the number of stimuli in a class (Fields & Verhave, 1987). No studies, however, have sought to identify the minimal mix of probes that would maximize the likelihood class formation or yield.

Although the "minimal mix" issue has not been addressed directly with MTS tests, it has been informed indirectly with sorting tests to document class formation (Arntzen, Granmo, & Fields, 2017; Arntzen, Norbom, & Fields, 2015b; Fields, Arntzen, & Moksness, 2014; Dickins, 2015; Pilgrim & Galizio, 1996; Smeets, Dymond, & Barnes-Holmes, 2000; Varelas & Fields, 2015). In some cases, after acquiring the baseline relations for three 5-member classes (A-B-CD-E), a sorting test was administered to assess class-emergence. This test began with the presentation of an initial stack of 15 cards, each corresponding to one class member. Producing three new piles that correspond to the three experimenter-defined classes indicated the immediate emergence of the classes.

The new piles would reflect control of behavior by a small number of stimulus relations for each class. In particular, assume that one card from each class is placed separately on table (e.g., A1—C2—E3). Placing the D1 stimulus on the A1 stimulus would reflect control of behavior by the D1–A1 relation, placing the B1 stimulus on the D1 stimulus would reflect control of behavior by the B1–D1 relation, placing the C1 stimulus on the B1 stimulus would reflect control of behavior by the C1–B1 relation, and placing the E1 stimulus on the C1 stimulus would reflect control of behavior by the E1–C1 relation, and likewise for the two other classes. Thus, sorting documents class formation with the administration of (N-1) probes per class. Outcomes such as these raise the possibility class formation could be documented with an MTS test that contained as few as (N-1) probes per class instead of the typical (N^2-N+1) probes per class.

It is important to note, however, that the results of sorting tests, at this point, should not be used as a new definition of equivalence classes. Also, the sorting data mentioned above were collected from typically functioning participants. More research will be needed to determine whether sorting tests can document class formation by individuals with developmental delays.

3. Yield does not take "near misses" into consideration

It could be argued that yield underestimates class formation because it does not consider "near misses," performances that are "close" to mastery and should be treated as indicating class formation. Near misses can be operationalized only when a quantitative value is used to define mastery. For example, if mastery is defined as a block of test trials in which at least 95% of the trials occasion class-indicative responses, near misses might be defined at accuracies of at least 90% but less than 95%, whereas accuracies less than 90% would not be considered near misses. Such an approach, however, does not provide an operational basis for making such a decision. If the near-miss accuracy is taken to indicate class formation (and it should not), the computation of yield would overestimate likelihood of class formation. If the near-miss accuracy is taken to indicate a lack of class formation (and it should), the computation of yield will underestimate the likelihood of class formation. The problem, however, cannot be solved by blaming yield as the culprit. Rather, near misses should be treated in a manner that permits a resolution of the ambiguity.

Although there is no current rationale for defining a near miss, we have adopted the following "halving strategy" to determine whether any performance that presumably represents a near-miss represents class formation. Assume that mastery is defined as a block that produces correct responses on at least 95% of the trials. Further, assume that a test block that produces an accuracy of 90% correct is judged to be a near miss. The trials in that block are divided into two halves with accuracies computed for each half. If accuracies in the first and second halves are 80% and 100% correct, respectively, that block reflected the rapid emergence of the classes; thus, the performance for the participant should be judged as showing class formation, and yield should include that participant as a "class-former." On the other hand, if accuracies in the first and second halves are 89% and 91% correct, respectively, that block would not reflect the emergence of an equivalence class. Thus, participants who generated those performances should not be included when computing yield.

Using this strategy with many of our experiments, we have found that the vast majority of test blocks that resulted in high but submastery performances showed class formation when their data were subjected to the halving procedure. On the other hand, when test blocks produced much lower submastery performances (e.g., 53%), the halving procedure rarely if ever showed the emergence of the classes. Indeed, this approach might be used with data provided by Vie and

Arntzen (2019). Disambiguating the outcomes of their near miss test blocks might change the yield-based outcomes of that experiment. On a broader level, if the outcomes of the halving procedure are obtained across many experiments, it should be possible to use that information to determine a data-based rationale for defining near misses, and enable yield to accurately reflect the effects of near misses on the likelihood of equivalence class formation.

4. Yield does not prompt use of sensitive statistical measures

Yield measures the prevalence of those who formed classes and excludes information for those who did not. Thus, it could be argued that yield is not as sensitive as measure of class formation because it does not include the data of all participants, and further. In addition, ANOVAs should be used for quantitative analysis instead of chi square or Fisher Exact tests. The validity of these strategies is questionable for two reasons.

First, the inclusion of data for all participants in a group combines performances by those who did and did not form classes. In virtually all of our experiments, those who formed classes produced far higher levels of accuracy than did those who did not form classes (e.g., ~100% and ~ 50%). Thus, combining data for these two "subgroups" would produce bimodal distributions of accuracy. The use of an ANOVA, whether parametric or nonparametric, necessarily assumes a unimodal population of scores. Because the empirical data sets are bimodal, the use of an ANOVA would be inappropriate, and indeed, noninformative. Thus, it is appropriate to use chi square or Fisher Exact tests to assess the significance of the effects of yield on class formation.

Second, the combining of data from these two subgroups would produce averages that do not reflect the performances of those who form classes or those who do not. Thus, on a descriptive level the averaged performances would provide misleading conclusions regarding the effect of a designated variable on class formation or its failure. For both reasons, the use of data averaged in this manner would be inappropriate.

It could also be argued that yield does not reflect the effects of sample size on the interpretation of outcomes. This issue can be considered from two perspectives. First, two experiments might produce the same yield, but might have been obtained from groups that contained different number of participants. Although nominally of equal yield, the two findings are not of equal power. In particular, the power of the effect in each condition is related to sample size. This can be quantified by computation of the binomial probabilities of obtaining each outcome, where the binomial probability of obtaining a given yield will increase sample size.

The other option is to consider different outcomes produced by different procedures. For example, assume that two conditions produced 30% and 70% yields. The inferences that can be drawn from these differences will vary based on the number of participants used to obtain each measure of yield. In particular, if it is assumed that both yields are based on sample sizes of 10, the difference in yields will not be significant. On the other hand, if the sample sizes for both yields are based on sample sizes of 100, the difference in yield will be significant. If, the 30% yield is based on a sample size of 10 and the 70% yield is based on a sample size of 100, the difference will also be significant.

These distinctions have little to do with the use of yield per se to make between-group comparisons in the same experiment, or between-group comparisons across experiments. Rather, this analysis points to the importance of being cognizant of the matching of sample sizes before making between-group comparisons of experimental outcome.

5. Yield does not identify variables that influence individual performances

Will a variable that produces a systematic increase in yield have a similar effect on class formation by individual participants? That question could be answered by conducting an experiment like that one conducted by Travis et al. (2014), who found that only 15% formed three-node, five-member A➔B➔C➔D➔E equivalence classes, whereas 85% did not form the classes. This yield was obtained in the absence of preclass-formation training. Yield, however, increased to about 80% when class-formation training was preceded by the establishment of simple discriminations using the C stimuli followed by 500 trials of overtraining. Thus, the likelihood of forming classes in a group increased by 65% (from 15% to 80%).

Would this group-based finding have similar effects on individuals? An answer to this question could be obtained by using the 85% of the participants who did not form classes in Travis et al.'s (2014) control group. These participants could be trained to establish C-based discriminations followed by overtraining as per Travis et al., and finally by the readministration of the ABCDE test blocks. If classes are formed by ~80% of those who did not initially form the classes, the same procedure would produce the same degree of enhanced class formation on with individuals and groups. Experiments such as these would show how yield-based outcome in group-based experiments would inform procedures to enhance class formation on an individual basis.

A large body of literature has shown that the percentage of participants who form equivalence classes can be substantially increased by the prior establishment of a stimulus control function with one of the potential class members (Arntzen & Nartey, 2018; Arntzen, Nartey et al., 2014; Arntzen, Nartey et al. 2015a; Arntzen, Nartey, & Fields, 2014b 2018a, b; Fields, Arntzen, Nartey, & Eilifsen, 2012; Nartey, Arntzen, & Fields, 2014; Nedelcu et al., 2015). Thus, could the degree of enhancement seen in these group studies also have similar effects on individual participants? This question could be answered by a strategy like that described in the preceding paragraph. For example, Nedelcu et al. (2015) found that ABCDE classes were formed by 77% of participants after the establishment of CV, CW, CX, CY, and CZ relations, i.e., C-VWXYZ training. In contrast, the ABCDE classes were formed by only 17% of the participants in a control group who received class-formation training with no preliminary C-VWXYZ training. To determine whether C-VWXYZ would have the same effect on single subjects, the 83% of the participants in the control group who did not form classes would be given C-VWXYZ training after failed class formation, which would be followed by a retest for the emergence of the ABCDE classes. If 77% of the 83% of participants form the ABCDE classes, that outcome would show that the C-VWZYZ would have the same effect on yields produced in a group and on a within subject basis. Thus, the yield-based group outcome would inform the effect of a variable on the formation of equivalence classes by individual participants.

6. Yield does not identify stimulus control during class formation

It could be argued that yield should not be used to index class formation because it cannot identify the many forms of stimulus control that are the determinants of responding during the delayed emergence of equivalence classes. Before considering the logical soundness of this assertion, the following is a brief summary of stimulus control factors that influence responding both during the delayed emergence of equivalence classes and after class formation.

Delayed emergence: accuracy-based measures One measure of delayed emergence is accuracy of responding (i.e., percentage of class indicative responding to trials of the same type in a block or session). Kennedy (1991) explored the gradual emergence of one-, second-, third-, four-, and five-node derived relations during the formation of five-node, seven-member equivalence classes. Early in testing, accuracy was an inverse function of the number of nodes that characterized the derived relations, and those relations then reached mastery in a temporal order that was a direct function of the nodal number of each type of derived relation. Fields, Adams, Verhave, and Newman (1993) studied delayed emergence during the expansion of class size. After forming three-member classes with A➔B➔C structures, C➔D relations were trained, and expansion of class size was assessed with the presentation of one-node DB and BD relations and two-node DA and AD relations. At the start of testing, the percentage of trials that produced class-consistent responding (i.e., accuracy) was greater with the one-node relations than the two-node relations. In addition, participants reached mastery of the one-node relations in fewer test blocks than the two-node relations. Similar results were reported by Kennedy, Itkonen, and Indquist (1994), who studied the delayed emergence of two-node, four-member classes having A➔B➔C➔D structures instead of class expansion. Most recently, Arntzen and Mensah (2020) found that delayed emergence was also influenced by the inclusion of a meaningful stimuli as the middle stimulus in a 3-node five member equivalence class, represented as A➔B➔Cm➔D➔E. Finally, Bentall, Jones, and Dickins (1998) showed reaction time was a direct function of nodal number early in testing and then became shorter and essentially constant with the continued testing that was administered during the delayed emergence of the classes. As noted above, however, one could argue that reaction time does not measure behavior, but rather is used as a proxy for cognitive activity (i.e., behavior that occurs during the presentation of the sample and comparison stimuli in the MTS trials).

Delayed or failed emergence: stimulus control topographies When intermediate levels of accuracy occurred, more detailed analyses of responding showed that behavior was controlled by a variety of relations amongst the stimuli, each called a stimulus control topography (SCT). These SCTs were identified by use of a *matrix* analysis that can measure a number of SCTs (Sidman, 1980; Sidman, Willson-Morris, & Kirk, 1986) or a *kernel* analysis that can measure up to 16 SCTs (Fields, Garruto, & Watanabe, 2010). Both sorts of analyses permitted the identification of control by (1) the location of one of the comparisons rather than the particular comparison in that location (i.e., POS); (2) a particular comparison stimulus regardless of its location among the locations of the comparison stimuli (i.e., COMP); (3) a given location based on the prevailing sample stimulus regardless of the comparison stimuli or their locations (i.e., SAMP); (4) the conditional relation between the sample and comparison from the same class (COND-DISC); or (5) a conditional relation between the sample and comparison but from different classes, e.g., participant-defined conditional discriminations (PD COND-DISC).

For example, Sidman (1980, 1992) and Iversen (2013) showed that during the acquisition of conditional discriminations, early in training, responding was determined by positionally defined stimulus control topographies, or by stimulus-control topographies in which preferences for a particular comparison stimulus was the determinant of responding. With continued training, these SCTs were replaced by the experimenter specified stimulus control topography.

Using the kernel analysis, Fields et al. (2010) showed control by upwards of four different nonclass-indicative stimulus control topographies during the delayed emergence of equivalence classes, the latter being documented by a fifth SCT: the class-indicative stimulus control topography.

The previously mentioned studies demonstrated the variety of stimulus control topographies that influenced responding during the delayed emergence of equivalence classes. On other occasions, participants did not form classes even with the repeated presentation of derived relations probes. When that occurred, responding does not necessarily reflect the random selection of comparison stimuli. In many instances, participants selected comparisons that reflected control of behavior by participant-defined relations (Arntzen, Nartey et al., 2015a; Mensah & Arntzen, 2017). For example, when presented with the XY probes for three potential classes, rather than conditionally selecting Y1 given X1, Y2 given X2, and Y3-given X3 (experimenter-defined relations), participants select Y1 given X3, Y2 given X1, and Y3 given X2 (participant-defined relations.). Thus, during delayed emergence or with failed class formation, probes performances reflected the control of behavior by participant-defined relations.

Accuracy, SCTs, and yield The experiments in this section documented the fact that nonclass-indicative stimulus-control topographies can influence responding during the acquisition of baseline relations, the delayed emergence of equivalence classes, and failed class formation. None of these findings, however, invalidate the use of yield to measure the formation of equivalence class that emerge on an immediate basis. Variables that influence delayed emergence can be measured with accuracy and SCTs but not with yield. On the other hand, variables that influence immediate emergence can be measured by yield but not by accuracy or SCTs. Thus, the combination of yield as well as trial-based measures of accuracy, SCTs, and response speed are needed for a comprehensive understanding of the variables that influence the equivalence class formation.

Summary and Conclusions

Yield has identified many variables that influence the immediate emergence of equivalence classes, and also has quantified the efficacy of equivalence-based instruction for teaching college-level academic content. In this article, we have listed six factors that could raise concerns about the validity of using yield to study equivalence class formation. Upon analysis, however, none support such a conclusion. Rather, they disclosed many nonyield-based factors that could influence the likelihood of class formation. To date, however, none have been explored experimentally. Rather, these analyses suggest new lines of research that could illuminate basic processes that influence the formation of equivalence classes.

Compliance with ethical standards

Ethical Approval All procedures performed in this study involving human participants, including informed consent, were in accordance with the ethical standards of the institutional research boards of the sited at which the cited experiments were performed, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of Interest Statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of Data and Materials Data can be found by contacting the first or second authors.

References

Adams, B. J., Fields, L., & Verhave, T. (1993). Effects of test order on intersubject variability during equivalence formation. *The Psychological Record, 43*, 133–152 Retrieved from http://thepsychologicalrecord.siuc.edu/index.html.

Albright, L. K., Fields, L., Reeve, K. F., Reeve, S. A., & Kisamore, A. N. (2019). Relatedness of equivalence class members: combined effects of nodality and relational type. *The Psychological Record, 69*, 277–289.

Arntzen, E. (2012). Training and testing parameters in formation of stimulus equivalence: methodological issues. *European Journal of Behavior Analysis, 13*, 123–135. https://doi.org/10.1080/15021149.2012.11434412.

Arntzen, E., & Holth, P. (2000). Equivalence outcome in single subjects as a function of training structure. *The Psychological Record, 50*, 603–628 Retrieved from http://thepsychologicalrecord.siuc.edu/index.html.

Arntzen, E., & Nartey, R. K. (2018). Equivalence class formation as a function of preliminary training with pictorial stimuli. *Journal of the Experimental Analysis of Behavior, 110*, 275–291. https://doi.org/10.1002/jeab.466 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/30182379.

Arntzen, E., Halstadtro, L. B., Bjerke, E., & Halstadtro, M. (2010). Training and testing theoretical music skills in a boy with autism using a matching-to-sample format. *Behavioral Interventions, 25*, 129–143. https://doi.org/10.1002/bin.301.

Arntzen, E., Halstadtro, L. B., Bjerke, E., Wittner, K. J., & Kristiansen, A. (2014a). On the sequential and concurrent presentation of trials establishing prerequisites for emergent relations. *The Behavior Analyst Today, 14*, 23–30. https://doi.org/10.1037/h0101280.

Arntzen, E., & Mensah, J. (2020). On the effectiveness of including meaningful pictures in the formation of equivalence classes. *Journal of the Experimental Analysis of Behavior*. https://doi.org/10.1002/jeab.579.

Arntzen, E., Nartey, R. K., & Fields, L. (2014b). Identity and delay functions of meaningful stimuli: Enhanced equivalence class formation. *The Psychological Record, 64*, 349–360. https://doi.org/10.1007/s40732-014-0066-3.

Arntzen, E., Nartey, R. K., & Fields, L. (2015a). Enhanced equivalence class formation by the delay and relational functions of meaningful stimuli. *Journal of the Experimental Analysis of Behavior, 103*, 524–541. https://doi.org/10.1002/jeab.152.

Arntzen, E., Norbom, A., & Fields, L. (2015b). Sorting: An alternative measure of class formation? *The Psychological Record, 65*, 615–625. https://doi.org/10.1007/s40732-015-0132-5.

Arntzen, E., Granmo, S., & Fields, L. (2017). The relation between sorting tests and matching-to-sample tests in the formation of equivalence classes. *The Psychological Record, 67*, 81–96. https://doi.org/10.1007/s40732-016-0209-9.

Arntzen, E., Nartey, R. K., & Fields, L. (2018a). Graded delay, enhanced equivalence class formation, and meaning. *The Psychological Record*. https://doi.org/10.1007/s40732-018-0271-6.

Arntzen, E., Nartey, R. K., & Fields, L. (2018b). Reorganization of equivalence classes: Effects of preliminary training and meaningful stimuli. *Journal of the Experimental Analysis of Behavior, 91*, 564–586.

Bentall, R. P., Jones, R. M., & Dickins, D. W. (1998). Errors and response latencies as a function of nodal distance in 5-member equivalence classes. *The Psychological Record, 49*, 93–115 Retrieved from http://thepsychologicalrecord.siuc.edu/index.html.

Bortoloti, R., & de Rose, J. C. (2009). Assessment of the relatedness of equivalent stimuli through a semantic differential. *The Psychological Record, 59*, 563–590 Retrieved from http://thepsychologicalrecord.siuc.edu/index.html.

Bortoloti, R., Rodrigues, N. C., Cortez, M. D., Pimentel, N. D. S., & de Rose, J. C. (2013). Overtraining increases the strength of equivalence relations. *Psychology & Neuroscience, 6*, 357–364. https://doi.org/10.3922/j.psns.2013.3.13.

Buffington, D. M., Fields, L., & Adams, B. J. (1997). Enhancing the formation of equivalence classes by pretraining of other equivalence classes. *The Psychological Record, 47*, 1–20.

Cullinan, V., Barnes, D., & Smeets, P. M. (1998). A precursor to the relational evaluation procedures: Analyzing stimulus equivalence. *The Psychological Record, 48*, 121–145.

De Souza, A. A., & Rehfeldt, R. A. (2013). Effects of dictation-taking and match-to-sample training on listing and spelling responses in adults with intellectual disabilities. *Journal of Applied Behavior Analysis, 46*, 792–804. https://doi.org/10.1002/jaba.75.

Devaney, J. M., Hayes, S. C., & Nelson, R. O. (1986). Equivalence class formation in language-able and language disabled children. *Journal of the Experimental Analysis of Behavior., 46*, 243–257.

Dickins, D. (2015). A simpler route to stimulus equivalence? A replication and further exploration of a “simple discrimination training procedure” (Canovas, Debert and Pilgrim 2014). *The Psychological Record*, 1–11. https://doi.org/10.1007/s40732-015-0134-3.

Doran, E. (2009). Analysis of variables manipulated in equivalence class research. (Unpublished Area Paper submitted to the Graduate School of CUNY).

Doran, E., & Fields, L. (2012). All stimuli are equal, but some are more equal than others: Measuring relational preferences within an equivalence class. *Journal of the Experimental Analysis of Behavior, 98*, 243–256. https://doi.org/10.1901/jeab.2012.98-243.

Dougher, M. (1994). Stimulus equivalence, functional equivalence, and the transfer of function. In S. C. Hayes, L. J. Hayes, M. Sato, & K. Ono (Eds.), *Behavioral analysis of language and cognition* (pp. 71–90). Reno, NV: Context Press.

Eikeseth, S., & Smith, T. (1992). The development of functional and equivalence classes in high-functioning autistic children: The role of naming. *Journal of the Experimental Analysis of Behavior, 58*, 123–134.

Fields, L., & Arntzen, E. (2018). Meaningful stimuli and the enhancement of equivalence class formation. *Perspectives on Behavior Science, 41*, 63–93. https://doi.org/10.1007/s40614-017-0134-5.

Fields, L., & Verhave, T. (1987). The structure of equivalence classes. *Journal of the Experimental Analysis of Behavior, 48*, 317–332. https://doi.org/10.1901/jeab.1987.48-317.

Fields, L., & Watanabe-Rose, M. (2008). Nodal structure and the partitioning of equivalence classes. *Journal of the Experimental Analysis of Behavior, 89*, 359–382.

Fields, L., Adams, B. J., Verhave, T., & Newman, S. (1993). Are stimuli in equivalence classes equally related to each other? *The Psychological Record, 43*, 85–105 Retrieved from http://thepsychologicalrecord.siuc.edu/.

Fields, L., Landon-Jimenez, D. V., Buffington, D. M., & Adams, B. J. (1995). Maintained nodal-distance effects in equivalence classes. *Journal of the Experimental Analysis of Behavior, 64*, 129–145.

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

Fields, L., Hobbie-Reeve, S. A., Adams, B. J., & Reeve, K. F. (1999). Effects of training directionality and class size on equivalence class formation by adults. *The Psychological Record, 49*, 703–724 Retrieved from http://thepsychologicalrecord.siuc.edu/.

Fields, L., Varelas, A., Reeve, K. F., Belanich, J., Wadhwa, P., DeRosse, P., et al. (2000). Effects of prior conditional discrimination training, symmetry, transitivity, and equivalence testing on the emergence of new equivalence classes. *The Psychological Record, 50*, 443–466 Retrieved from http://thepsychologicalrecord.siuc.edu/index.html.

Fields, L., Travis, R., Roy, D., Yadlovker, E., de Auguiar-Rocha, L., & Sturmey, P. (2009). Equivalence formation: A method for teaching statistical interaction. *Journal of Applied Behavior Analysis, 42*, 575–593. https://doi.org/10.1901/jaba.2009.42-575.

Fields, L., Garruto, M., & Watanabe, M. (2010). Varieties of stimulus control in matching-to-sample: A kernel analysis. *The Psychological Record, 60*, 3–26 Retrieved from http://thepsychologicalrecord.siuc.edu/.

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*, 163–181. https://doi.org/10.1901/jeab.2012.97-163.

Fields, L., Arntzen, E., & Moksness, M. (2014). Stimulus sorting: A quick and sensitive index of equivalence class formation. *The Psychological Record, 64*, 487–498. https://doi.org/10.1007/s40732-014-0034-y.

Fienup, D. M., Wright, N. A., & Fields, L. (2015). Optimizing equivalence based instruction: Effects of training protocols on equivalence class formation. *Journal of Applied Behavior Analysis, 48*, 1–19. https://doi.org/10.1002/jaba.234.

Goyos, C. (2000). Equivalence class formation via common reinforcers among preschool children. *The Psychological Record, 50*, 629–654.

Iversen, I. H. (2013). Matching-to-sample performance in rats: a case of mistaken identity? *Journal of the Experimental Analysis of Behavior, 68*, 27–45.

Kennedy, C. H. (1991). Equivalence class formation influenced by the number of nodes separating stimuli. *Behavioural Processes, 24*, 219–245. https://doi.org/10.1016/0376-6357(91)90077-D.

Kennedy, C. H., Itkonen, T., & Indquist, K. (1994). Nodality effects during equivalence class formation: An extension to sight-word reading and concept development. *Journal of Applied Behavior Analysis, 27*, 673–683. https://doi.org/10.1901/jaba.1994.27-673.

Lazar, R. M. (1977). Extending sequence-class membership with matching to sample. *Journal of the Experimental Analysis of Behavior, 27*, 381–392.

Mensah, J., & Arntzen, E. (2017). Effects of meaningful stimuli contained in different numbers of classes on equivalence class formation. *The Psychological Record, 67*, 325–336. https://doi.org/10.1007/s40732-016-0215-y.

Moss-Lourenco, P., & Fields, L. (2011). Nodal structure and stimulus relatedness in equivalence classes: post-class formation preference tests. *Journal of the Experimental Analysis of Behavior, 95*, 343–368. https://doi.org/10.1901/jeab.2011.95-343.

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

Nartey, R. K., Arntzen, E., & Fields, L. (2015a). Enhancement of equivalence class formation by pretraining discriminative functions. *Learning & Behavior, 43*, 20–31. https://doi.org/10.3758/s13420-014-0158-6.

Nartey, R. K., Arntzen, E., & Fields, L. (2015b). Training order and structural location of meaningful stimuli: Effects on equivalence class formation. *Learning & Behavior, 43*, 342–353. https://doi.org/10.3758/s13420-015-0183-0.

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

Ong, T., Normand, M. P., & Schenk, M. J. (2018). Using equivalence-based instruction to teach college students to identify logical fallacies. *Behavioral Interventions.* https://doi.org/10.1002/bin.1512.

Pilgrim, C., & Galizio, M. (1996). Stimulus equivalence: A class of correlations or a correlation of classes?. In T. R. Zentall & P. M. Smeets (Eds.) *Stimulus Class Formation in Humans and Animals*. New York, NY: Elsevier Science B. V.

Saunders, K. J., Saunders, R. R., Williams, D. C., & Spradlin, J. E. (1993). An interaction of instructions and training design on stimulus class formation: extending the analysis of equivalence. *The Psychological Record, 43*, 725–744.

Saunders, R. R., Chaney, L., & Marquis, J. G. (2005). Equivalence class establishment with two-, three-, and four-choice matching to sample by senior citizens. *The Psychological Record, 55*, 539–559.

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

Sidman, M. (1980). A note on the measurement of conditional discrimination. *Journal of the Experimental Analysis of Behavior, 33*, 285–289. https://doi.org/10.1901/jeab.1980.33-285.

Sidman, M. (1987). Two choices are not enough. *Behavior Analysis, 22*, 11–18.

Sidman, M. (1992). Adventitious control by the location of comparison stimuli in conditional discriminations. *Journal of the Experimental Analysis of Behavior, 58*, 173–182.

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*, 127–146. https://doi.org/10.1901/jeab.2000.74-127.

Sidman, M., & Cresson, O. (1973). Reading and crossmodal transfer of stimulus equivalence in severe retardation. *American Journal on Mental Retardation, 77*, 515–523 Retrieved from http://aaidd.org/publications/journals#.UsAc2PZZWX0.

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.

Sidman, M., Willson-Morris, M., & Kirk, B. (1986). Matching-to-sample procedures and the development of equivalence relations: The role of naming. *Analysis & Intervention in Developmental Disabilities, 6*, 1–29. https://doi.org/10.1016/0270-4684(86)90003-0.

Slotnick, B. M., & Silberberg, A. M. (1993). Odor matching and odor memory in the rat. *Physiology and Behavior, 53*, 795–804.

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.

Spear, J., & Fields, L. (2015). Effects of two parameters of joint stimulus control training on the induction of expressive writing: Learning to write without writing. *Learning and Behavior, 43*, 354–375. https://doi.org/10.3758/s13420-015-0184-z. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26077442.

Spencer, T. J., & Chase, P. N. (1996). Speed analysis of stimulus equivalence. *Journal of the Experimental Analysis of Behavior, 65*, 643–659.

Steele, D. L., & Hayes, S. C. (1991). Stimulus equivalence and arbitrarily applicable relational responding. *Journal of the Experimental Analysis of Behavior, 56*, 519–555.

Taylor, I., & O'Reilly, M. F. (2000). Generalization of supermarket shopping skills for individuals with mild intellectual disabilities using stimulus equivalence training. *The Psychological Record, 50*, 49–62.

Tomanari, G. Y., Sidman, M., Rubio, A. R., & Dube, W. V. (2006). Equivalence classes with requirements for short latencies. *Journal of the Experimental Analysis of Behavior, 85*, 349–369.

Travis, R. W., Fields, L., & Arntzen, E. (2014). Discriminative functions and over-training as class-enhancing determinants of meaningful stimuli. *Journal of the Experimental Analysis of Behavior, 102*, 47–65. https://doi.org/10.1002/jeab.91.

Varelas, A., & Fields, L. (2015). Induction of a generalized transitivity repertoire via multiple-exemplar training and staged testing. *The Psychological Record, 65*, 595–614. https://doi.org/10.1007/s40732-015-0129-0.

Vie, A., & Arntzen, E. (2019). Role of distractors in delayed matching-to-sample arrangements in tests for emergent relations. *International Journal of Psychology & Psychological Therapy, 19*, 71–88.

Walker, B. D., & Rehfeldt, R. A. (2012). An evaluation of the stimulus equivalence paradigm to teach single-subject design to distance education students via Blackboard. *Journal of Applied Behavior Analysis, 45*, 329–344. https://doi.org/10.1901/jaba.2012.45-329.

Wang, T., Dack, C., McHugh, L., & Whelan, R. (2011). Preserved nodal number effects under equal reinforcement. *Learning & Behavior, 39*, 224–238. https://doi.org/10.3758/s13420-011-0020-z.

Wirth, O., & Chase, P. N. (2002). Stability of functional equivalence and stimulus equivalence: effects of baseline reversals. *Journal of the Experimental Analysis of Behavior, 77*, 29–47.

Wulfert, E., Dougher, M. J., & Greenway, D. E. (1991). Protocol analysis of the correspondence of verbal behavior and equivalence class formation. *Journal of the Experimental Analysis of Behavior, 56*, 489–504.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.