Journal of Behavioral Education
https://doi.org/10.1007/s10864-022-09485-2

ORIGINAL PAPER

Comparing the Detection Accuracy of Operational Definitions and Pinpoints

Richard M. Kubina Jr.[1] · Madeline Halkowski[1] · Kirsten K. L. Yurich[2] · Kimberly Ghorm[3] · Nora M. Healy[3]

Accepted: 4 July 2022

Abstract

Operational definitions have a significant history in applied behavior analysis. The practice's importance stems from the role operational definitions play in detecting an event, human thought, or action. While operationalizing target behaviors has enjoyed widespread practice, some concerns have recently arisen with translation validity and detection accuracy. Additionally, a review of the literature produces few articles assessing the validity of operational definitions. Pinpoints represent an alternative for describing target behaviors. A pinpoint has a formula for construction that includes using an action verb, an object, or event that receives the action, and a comprehensibly defined context where the observation of the action verb + object or event occurs. Pinpoints also have few empirical studies demonstrating their validity. The following experiment compared the detection accuracy of an operational definition for self-injurious behavior and a corresponding pinpoint across professionals who worked in a school that served clients with autism spectrum disorder. The results indicate lower accuracy scores for the operational definition when compared to the pinpoint. Additionally, the consistency of scores varied more for the operational definition than the pinpoint.

Keywords Operational definition · Pinpoint · Detection accuracy · Self-injurious behavior · Autism spectrum disorder

✉ Richard M. Kubina Jr.
RMK11@psu.edu

[1] Special Education Program, The Pennsylvania State University, 209 CEDAR Building, University Park, PA 16802-3109, USA

[2] Felician University, Rutherford, USA

[3] Vista Autism Services, Hershey, USA

Published online: 10 August 2022

Comparing the Detection Accuracy of Operational Definitions and Pinpoints

Behavior analysis has a well-earned moniker—the science of behavior. Four domains of behavior analysis include radical behaviorism, the experimental analysis of behavior, applied behavior analysis, and practice guided by behavior analysis (Cooper et al., 2020). The remarkable database of experimental, applied, and practice outcomes appears in a trove of behavior analytic, psychology, and other related journals. Because behavior analysis functions as a science, there remain many undiscovered functional relations. Also, the methods used within behavior analysis require continual examination, scrutinization, and possible change. Operational definitions represent one such practice in need of further inspection.

Recommendations for operational definitions in applied and practice settings occur in behavior analytic textbooks (e.g., Alberto et al., 2022; Cooper et al., 2020; Mayer et al., 2019) and single-case methods texts (e.g., Kazdin, 2020; Ledford & Gast, 2018). Textbooks champion definitions produce target behaviors with characteristics such as precision, accuracy, clarity, completeness, and concision. Yet an actionable, universal framework for producing high-quality operationalizations does not exist. For example, Alberto et al. (2022) shared eight different operational definitions of on-task behavior found in the several studies. The structure of all definitions varied along the following dimensions:

- A list of categories of specific behaviors
- An abbreviated list of representative behaviors
- Examples demonstrating the presence and absence of the behavior
- Inclusion of positive examples
- Inclusion of negative examples
- Adding a time element

Beyond textbooks, there exist almost no behavior analytic experiments examining the feasibility, technical adequacy, and reliability of operational definitions. However, one can find a long-standing debate and literature beginning in the 1930s surrounding the philosophy of operationalism. The scholarly exchange of ideas led to a famous symposium on the topic led by Skinner (1945) and featured major figures such as the philosopher Percy Bridgman (1945), logical empiricist Herbert Feigl (1945), and the famous psychologist Edwin Boring (1945). The symposium did not lead to an agreement among the many participants. The subsequent years following the symposium led to more discussions surrounding ideas concerning operationalism. Still, no agreed-upon empirically validated method for producing coherent, internally consistent definitions of behavior appeared in the literature.

In the psychology literature, operational definitions have received criticism on theoretical and practical grounds. For example, translation validly refers to the "the closeness with which the study's intended meaning of constructs matches their operationalization" (Krathwohl, 2009, p. 405). Translation validity has three relevant limitations (Slife et al., 2016). First, operational definitions differ from the constructs they represent. For example, "on-task" typifies a construct used in experiments in behavior analysis, psychology, and education experiments. On-task represents a student's focus or productivity. A study on a second-grade math class defined on-task as students commenting on achievements made in a game, reporting successes or failures during an exercise, or asking the teacher or another student for help with a learning task (Beserra et al., 2019). The researchers translated on-task into behaviors that they could observe and count. What the researchers selected may not represent student focus or productivity. Asking for help from a peer could occur for attention-seeking purposes. Commenting on their game progress might happen to boast about achievement or tease a peer. Translation validity came from the researchers' preference and pronouncement.

Second, any construct in need of operationalization by its nature does not rise to the standard of measurability. Therefore, operationalizing and creating a measurable event does nothing to validate the relationship between the construct and the selected measure (Slife et al., 2016). "Tolerance" shows how a construct falls outside of measurement parameters. Tolerance refers to situations where an individual must wait to access the desired event, wait for something they do not like to decrease, or someone withholding a preferred event (O'Rourke et al., 2019). Stated differently, to tolerate an unpleasant situation or event, the individual must successfully refrain from any behavior that indicates intolerance. Tolerance becomes apparent with the absence of an action in prescribed circumstances. Therefore, no direct measures exist which can support the translation validity of the construct.

Third, due to the lack of direct measurement between the construct and operationalization, having multiple instances of an operational definition do not yield translation validity (Slife et al., 2016). The counter to on-task, off-task, illustrates the issue. A large-scale study examining elementary school students' attention allocation during instructional activities classified off-task as instances when students did not look at the teacher, instructional materials, or instructional activity (Godwin et al., 2016). Another study evaluated noncontingent reinforcement and defined off-task during class time as calling out, chatting with peers, not looking at the teacher or task, drawing items or coloring outside of the assigned task, or leaving one's seat (Austin & Soeda, 2008). A third study examined a class-wide positive behavior support program and identified off-task as off-task motor, off-task verbal, or off-task passive (Kraemer et al., 2012). The multiple instances of operationalization do not converge, nor do they reveal an underlying identity (i.e., off-taskness) establishing construct validity.

Beyond the questionable relationship between operationalization and translation validity, a research study employing an operational definition raised concerns between reliable detection and measurement of behaviors targeted for change. Smith et al. (2013) examined participants' accuracy in detecting a student's target behavior when identifying the behavior using two different forms of a target behavioral definition (TBD). The first definition (TBD1) followed an operational definition used by the student's school. The second definition (TBD2) featured a shorter phrase that included only an object term and present tense, active verbs describing movements associated with the head hits. The results demonstrated participants' average percent accuracy for TBD1 came to 35%, while TBD2 had 68% accuracy (Smith et al., 2013). The researchers postulated that the difference in participant performance stemmed from the inclusion of additional descriptive words such as "swings forcefully" in TBD1, which often appear in operational definitions to produce a more comprehensive picture of the behavior. The previously mentioned phrases can unintentionally increase the overall subjectivity of a TBD and can negatively affect detection accuracy.

Operational definitions have five overall shortcomings from a theoretical, practical, and applied perspective. First, when used, the inability to measure a construct produces translation validity issues. Second, no agreed upon method for producing reliable, valid operational definitions has emerged in the research literature, behavior analytic textbooks, or psychology research literature or textbooks. Third, operationalizing involves linking selected behaviors back to the original target behavior. The variability produced by individual researchers or research teams leads to inconsistency and multitudinous actions that may or may not correspond to a common description of a behavior. Fourth, scant empirical knowledge exists supporting operational definitions as the ideal method for defining target behaviors. And fifth, emerging evidence demonstrates the use of operational definitions can lead to low signal detection for measuring a target behavior (Smith et al., 2013), providing questionable or inaccurate descriptions of clinical targets (Breitborde et al., 2009; Spira et al., 2015), and does not match with the actual events of people who experienced the operationalized target behavior (Menin et al., 2021). One alternative to operationalization comes from the precision teaching literature, a pinpoint.

Precision Teaching and Pinpointing

Precision teaching (PT) began in the late 1960s as a system to help parents and teachers precisely measure behavior, analyze data, and make decisions about a child or student. The founder of PT, Ogden Lindsley, hoped parents and teachers could implement lessons learned from Skinner and successful applications of behavior analysis (Lindsley, 1972, 1990, 1991). PT has four steps: pinpoint, record, change, and try again. The first step of pinpointing actions or thoughts anchors the PT process (Kubina, 2019). Without the ability to detect a target, collecting data, analyzing it, and trying additional problem-solving interventions to improve behavior will yield questionable results.

The unique framework for creating pinpoints evolved as PT matured. In the 1970s and 1980s, the term "movement cycle" referred to a behavior that a teacher or parent could directly observe, had a clear beginning and end, and constituted a cycle or could repeat itself (White & Haring, 1980). For example, instead of operationally defining "aggression," the movement cycle would focus on a precise action. "Kicks leg" or "punches peer" would serve as the movement cycle. Both "kicks leg" and "punches peer" allow for clear observation, have explicit beginnings and endings for each instance of the action, and are repeatable. Furthermore, the movement cycle aligned with Skinner's movement-based definition of behavior: Behavior is what an organism is doing-or more accurately what it is observed by another organism to be doing" (Skinner, 1938, p. 6).

As PT advanced, the term movement cycle changed to pinpoint to reflect the addition of context (Kubina & Yurich, 2012). For instance, "writes name" described an action clearly but lacked the context of where, when, with whom, or with what the movement cycle occurred. "Writes name *on chalkboard*" and "writes name *with pen*" or "writes name *in the sand*" all reflect different behaviors. A fully formed pinpoint enhances detection of the target behavior, directly represents the target selected for observation, and improves communication among stakeholders (Kubina, 2019).

Comparing pinpoints with operational definitions offers two interesting contrasts. First, a properly formed pinpoint would not have translation validity issues due to its explicit construction. Each pinpoint has three components expressed with words depicting observable behavior: (1) an unambiguous action verb, (2) an equally distinct object or event that receives the action (i.e., the first two components form a movement cycle), and (3) the clearly defined context in which the movement cycle takes place (e.g., bites fingernail during a chemistry test). Instead, a construct like anxiety may require multiple different behaviors and subjective examples; a pinpoint specifies the exact target an observer would detect and count without further definitions.

Second, pinpoints may also provide better detection accuracy than operational definitions (Smith et al., 2013). While pinpoints have a long history of applied use, we could find no studies that empirically assessed their technical adequacy. Therefore, examining if a practical difference occurs between pinpoints and operational definitions would shine a light on each method for labeling target behaviors.

Schools for students with disabilities and learning differences serve as one area of practice acutely in need of superior detection accuracy. The following study took place in a school for students with autism spectrum disorder. For students who exhibit challenging behavior, staff must detect a targeted behavior and accurately count its occurrence. The data generated from each count become the data behavior analysts and other clinicians use to guide their treatment decisions. At the time of this writing, we could find no other study examining the difference between operational definitions and pinpoints with respect to detection accuracy. Therefore, we asked the following two experimental questions. First, to what extent will an operational definition differ from a pinpoint in detection accuracy of videos depicting a young man engaging in self-injurious behavior, specifically head hitting? Second, how will participants' rates of false positives or false negatives compare depending on the type of behavior definition they use to identify the selected target behavior?

Method

Participants

Twenty-eight employees (21 female; 7 male) from a private school in Pennsylvania participated in the study. The school serves students aged 3 to 21 with the primary diagnoses of autism spectrum disorder. Participants held positions as behavior technicians ($n = 17$), personal care assistants ($n = 4$), educational behavior support employees ($n = 3$), employment specialists ($n = 2$), and pre-employment specialists ($n = 2$). The roles required them to complete ongoing training in behavior support practices and safety procedures as part of the school's commitment to professional development. The participants reported their highest level of education as bachelor's ($n = 26$) and master's ($n = 2$), which mirrors the education level of most registered behavior technicians in the field (Carr & Nosik, 2017; Novack & Dixon, 2019).

Recruitment procedures included random sampling via email invitations to each department and posting flyers across the main campus and satellite locations; no one received in-person solicitations. Selection criteria included a minimum of one year of employment history at the school, availability during regular business hours, and no restrictions on demographics. Once a staff member requested admission into the study, they met with the school's human resources department to complete the informed consent process. Twenty-eight chose to participate out of thirty individuals who expressed interest. The participants received an optional $15 Amazon gift card for their involvement in the study as an incentive.

Materials

Two video presentations served as the stimuli for the study. Each video contained 30, 10-s video segments separated by a 5-s black screen. The black screen displayed a 3-s count down to the next video segment. The videos depicted a young man engaging in self-injurious behavior across a variety of settings.

We created the video presentations by selecting seven different pre-recorded behavior episodes that the client's family had made publicly available on a digital media platform. Each video portrayed the individual engaging in the same target behavior over several years and across multiple settings. The seven primary source videos broke down into 10-s segments (60 segments in total). We then placed the segments in a quasi-random order to create a balanced ratio of target behavior occurrences across both videos. Video One contained 159 instances of behavior across 20 of the 30 video segments, with the average segment containing five instances of behavior (SD = 7). Video Two contained 145 instances of behavior across 18 of the 30 video segments, averaging five instances of behavior within each segment (SD = 6).

Participants viewed the stimuli on 15-inch Lenovo laptops provided by the organization's quality, training, and research department. The laptops sat on cafeteria-style tables, creating a viewing distance of approximately 2–3 feet from participants. Each screen's setting had full brightness levels with the sound turned off. Three-foot cardboard trifold dividers stood 18 inches tall, creating a border around each seat, which helped minimize visual distractions.

At the beginning of each session, participants received a black ballpoint pen, a data collection sheet, and a behavior definition card. The scoring sheet consisted of a 2-by-30 grid with column one listing each segment (1–30), while column two provided blank spaces for participants to mark their responses. Experimenter materials included a timer and a clipboard with copies of the attendance sheet and the procedural script.

Experimental Design

The study employed a two-period crossover design, as depicted in Fig. 1. Group 1 ($n = 14$) received the pinpoint first, followed by the operational definition, while Group 2 ($n = 14$) obtained the behavior descriptions in reverse order. We inserted a 24-h washout period (no treatment) between sessions one and two to help decrease the risk of a carryover effect (Sibbald & Roberts, 1998). Each participant also acted as their own control, allowing the study to achieve the same statistical power with fewer participants (Piantadosi, 2005; Putt & Chinchilli, 2004; Reed, 2004).

Independent Variable

Two target behaviors served as the independent variable. The first followed the format of an operational definition, described topographically (Cooper et al., 2020; Johnston et al., 2020; Mayer et al., 2019). Operational definitions adhere to general guidelines for best practice rather than specific structural criteria. Therefore, we reviewed the literature to find examples of acceptable operational definitions (see Table 1) and selected one that (a) included all forms of the behaviors shown in the video and (b) exemplified the guidelines for "clear, objective, and concise" (Cooper et al., 2020).

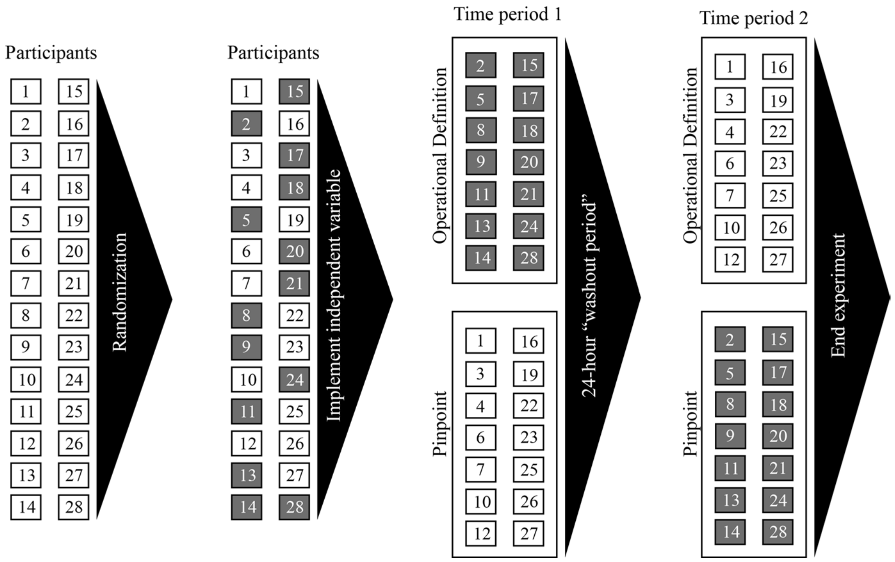

Fig. 1 A visual representation of the AB:BA design. *Note*: All initial white boxes represent unassigned participants. Subsequent gray and white boxes indicate randomized and then grouped participants

Table 1 Examples of operational definitions from the literature

References	Direct quote
Bird et al. (2020)	"Self-injury was defined as an episode during which Luther attempted to or actually struck knees against his head, hit his head against a fixed surface or object, punched his face and head with hands, pressed fingers against his eyes, and bit any part of his body" (p. 139)
Courtemanche et al. (2018)	"We defined head hitting as any instance in which the participant used an open palm, closed fist or object to make contact with any area of the head or face with sufficient force to (1) produce a sound audible via video or (2) move the head at least1in.We defined headbanging as any instance in which the participant's head (back, front or side) contacted a piece of furniture, door or wall with sufficient force to produce a sound audible via video." (p. 110)
Gerow et al. (2021)	"*Self-injurious behavior* was defined as behavior likely to cause harm to the child (e.g., hand-to-head hitting, biting fingers)" (p. 58)
Gregori et al. (2018)	"SIB is defined as actions that produce, or attempt to produce, physical damage to one's body (Tate & Baroff, 1966) " (p. 112)
Hagopian et al. (2015)	"Head-directed SIB, which included behaviors that targeted the individual's head area (e.g., punching self in face, head banging)" (pp. 526–227)
Huisman et al. (2018)	"[…] we propose to define SIB as non-accidental behavior resulting in demonstrable, self-inflicted physical injury, without intent of suicide or sexual arousal" (p. 484)
Iwata et al. (1982)	"Head Hitting: Forceful contact of the hand with any part of the head" (p. 8)
Linscheid et al. (1990)	"Head banging: any forcible contact between hand and head or between head and object" (p. 59)
Linscheid et al. (1994)	"Head hit: Any forceful contact by the hand directed at the head" (p. 84)
Robinson et al. (2019)	"Tony's educators defined his SIB as anytime he bangs his head against tables, walls, and doors, and anytime Tony slaps himself in the face with open hands and hits himself in the face with a closed fist" (p. 157)
Rooker et al. (2018)	" […] (1)'sharp' SIB if the contact between body parts included either nails or teeth (this included self-biting, pinching and scratching); or (2)'blunt' SIB, if the contact did not involve nails and teeth but was between other body parts or body parts and the environment (this included body-hitting, body to surface, self-kicking, head banging, head-hitting, object to head, shoulder to head and knee to head)" (p. 1089)
Shore et al. (1994)	"Self-injurious behaviors were defined as follows: *head of body hitting*—audible contact of a hand, fist, or knee against any part of the face, head, or body; *head banging*—contact of the head with a stationary object" (p. 373)
Smith et al. (1996)	"SIB observed during this study included head hitting (contact of palm or fist against head)" (p. 99)
Szymanski et al. (1987)	"Head hits: any part of one or both hands strikes any region of face, helmet or shield. Excluded are light touches or strokes with fingertips. 2. Head bangs: any part of head or helmet strikes against floor, furnishings, or other stationary object. May occur from any position (sitting, standing, etc.)" (p. 184)

The original definition read, "[head hitting is the use of] an open palm, closed fist, or object to make contact with any area of the head of face with sufficient force to (1) produce audible sound via video or (2) move the head at least 1 inch" (Courtemanche et al., 2018, p. 1100). However, the home video segments used in the compilation video emitted poor, uneven sound quality and the occasional vocal outbursts from people on and off the screen. Because the external sounds were not associated with the target behavior and could potentially distract viewers, we removed the sound from the video to standardize the viewing experience and avoid potential confounds. The absence of sound affected the auditory criteria in the original definition. The final version of the operational definition read as, "Head hitting: using an open palm, closed fist, or object to make contact with any area of the head or face with sufficient force to move the head at least one inch."

The second definition matched the format of a pinpoint, which consists of a movement cycle and a context (Kubina, 2019). A movement cycle includes an active verb, expressed in the simple present tense, and a word or short phrase to describe the "object receiving the action" (Kubina, 2019. p. 19). For example, the movement cycle "circles letter" describes the action taken (i.e., circle) and the object (i.e., letter). The context "specifies the where, when, with whom, or with what" (Kubina, 2019, p. 22). Some examples of context include *in* a jar, *when* the light stops, *with* a pencil. Based on the selected operational definition, the following pinpoint served as the second behavior definition: "hits head with hand or fist."

The behavior descriptions contained different criteria for behavior identification because the guidelines for practitioners' use differ significantly between operational definitions and pinpoints. Even though both descriptions seek clarity, they differ in their approach. Operational definitions pursue accuracy by including numerous examples explaining how the behavior may appear in different situations. In contrast, pinpoints attempt precision by simplifying the target behavior description to its most critical element. As researchers, we noticed the differences and sought to investigate how these two strategies of behavior definitions impact behavior technicians' ability to identify behavior. In short, the differences in independent variables occur by design.

Dependent Variable

The count of target behaviors recorded by participants under each behavior condition (i.e., operational definition, pinpoint) represent as the dependent variable. Participants viewed the video presentations during the experimental sessions, marking each perceived instance of the target behavior on a data collection sheet after a recorded segment. We analyzed responses using total count and segment-by-segment calculations.

Procedure

The study took place over two days, with experimental sessions held at 9:30 AM, 10:15 AM, and 11:00 AM. Each participant had to attend one of the three sessions to ensure an equal number of participants from Group 1 and Group 2 attended each time slot (Fig. 1). After the random assignment, we re-assigned five participants due to scheduling conflicts. Sessions ran approximately 30 min in length, followed by a fifteen-minute gap for sanitization of the environment. Participants received a data collection sheet, a pen, and a behavior definition card.

On day one, participants completed a demographic survey and a brief practice session. We included the practice session so researchers could confirm that all participants had the ability to correctly record frequency behavior using a pen and paper. During the practice session, participants watched a video depicting an adult male ripping pieces of paper in front of a small child who emitted bursts of laughter after each rip. We edited the video to follow the same structure as the experimental video: a 10-s video clip followed by a 3-s countdown displayed on a black screen (repeated ten times). The fourth author demonstrated how to create a mark on the data collection sheet each time the adult ripped a piece of paper. If no instance of target behavior occurred during a given segment, the researcher instructed the participants to write a "0." Participants then practiced marking the behaviors while the fourth and second authors observed and provided corrective feedback. Because participants collected behavioral data on iPads as part of their typical work at the school, the training sessions allowed the research team to ensure that the participant's data collection capabilities generalized to a pen and paper format. We did not assess the accuracy of the participants tallies but rather the manner in which they made them. Once participants demonstrated 100% accuracy, the fourth author instructed the participants to begin the experimental video. The participants could not pause, fast-forward, rewind, or stop the video during the actual study.

Participants turned their data collection sheet faced down when the video presentations ended. We provided verbal reminders not to discuss the study's contents outside the testing site. The participants stayed seated as we collected their datasheets and dismissed them one at a time. Day two followed the same procedures and scripts excluding the training session.

Gold Standard

A "gold standard" represents a reference for the best available measure of the presence or absence of a condition (Trikalinos et al., 2012). When practitioners record target behaviors in applied settings, the inability to *stop, pause,* or *rewind* the event allows for human error to affect accuracy. Therefore, to create the gold standard, we used the online software program, Vosaic, which allows researchers to analyze and annotate videos using frame-by-frame viewing (Ehrenfeld & Horn, 2020; Smith et al., 2013; Vosaic, 2020). Vosaic's software allows multiple people to annotate a single video while keeping their marks invisible from their peers. Five researchers (i.e., one BCBA-D, two BCBAs, and two researchers trained in behavior analysis) independently viewed the videos and marked each instance of self-injurious behavior as defined by the behavior descriptions, adhering to variations in verbiage. Once each researcher completed annotating the videos individually, they met to compare results and discuss any areas of disagreement. The gold standard emerged when observers achieved 100% agreement on all 60 video segments.

The researchers used both behavior definitions when determining the gold standard, carefully considering the differences between each description. The operational definition, for instance, included phrases such as, "[uses] object to make contact with any area of the head or face" and "with sufficient force to move the head at least one inch." However, the inclusion of these phrases did not result in differences in the gold standard. In several video segments, the individual engages in self-injurious behavior with his hands wrapped in blankets. The occurrences counted as instances of behavior under the operational definition criteria "[uses] object to make contact" as well as for the pinpoint "hits head with hand of fist." The "one inch" criteria also did not create a disparity between the descriptions because each time the individual hit their head, it moved about one inch.

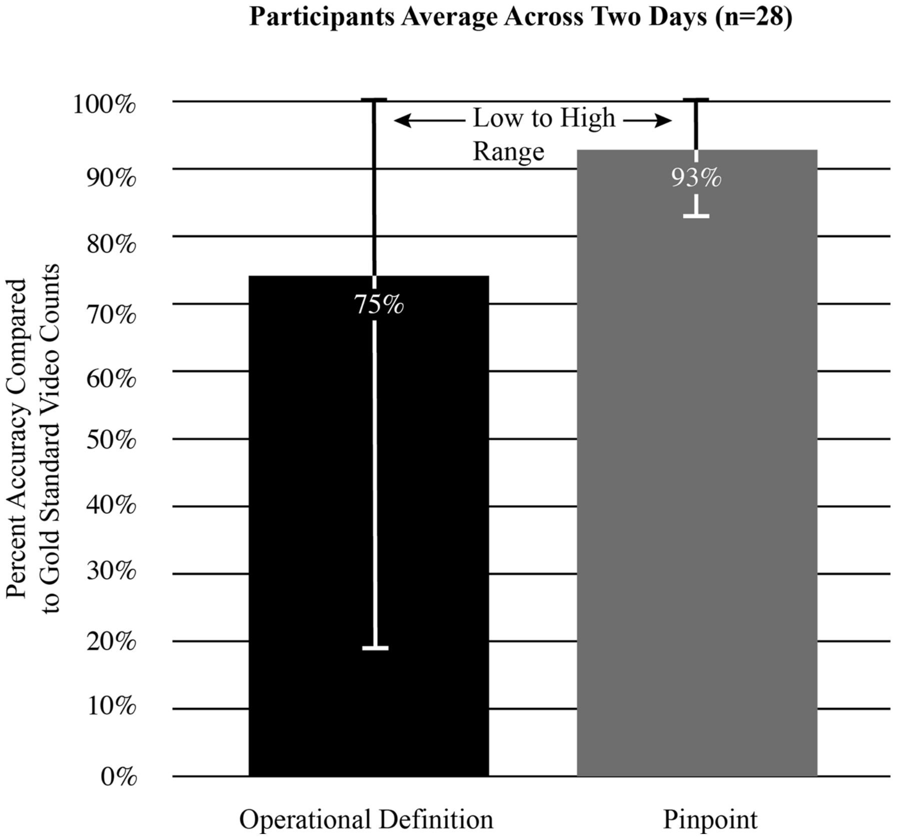

Fig. 2 The overall accuracy of participants' accuracy under operational definition and pinpoint conditions

Data Analysis

We calculated percent agreement by comparing the total count of target behaviors recorded by participants under each condition (i.e., operational definition and pinpoint) to the count of target behaviors identified in the gold standard (Smith et al., 2013). Figure 2 shows the mean percent agreement for both conditions. Furthermore, lines with horizontal dash endcaps illustrate the range of agreement from low to high. Only included segments that contained an instance of the target behavior as per the gold standard or segments that a participant indicated an instance of behavior had occurred factored into the percent agreement calculation.

An agreement score happened when a participant identified the target behavior's presence within a video segment (i.e., true positive), and the gold standard did as well. A disagreement occurred when the gold standard indicated an instance of the target behavior, but the participant did not report any (i.e., a false negative). Another example of a disagreement occurred during a segment with zero instances of the target behavior, but a participant reported an instance occurred (i.e., a false positive). We determined percent agreement by dividing the total number of agreements by the sum of agreement and disagreements for each participant in each condition (Smith et al., 2013).

Procedural Integrity

The study used procedural scripts to increase consistency across all sessions. The scripts contained the participants' directions regarding the initial intake process, prestudy training video, study rules, and directions for exiting the testing site. Before the first day of the study, we reviewed the script and role-played the study procedures until we could perform all tasks fluently. Then multiple observers addressed procedural integrity across the study by monitoring the fourth author as they read the script and directed participants on how to use tally marks to record behavior frequency on the data collection sheets. Two other researchers positioned themselves around the room to monitor participants' actions and the actions of the other researcher. Two researchers per sessions took data on procedural integrity by checking the verbiage used by the fourth author against a copy of the procedural script. Each session the instructor completed 100% of the procedural script.

Results

The study compared professionals' ability to accurately detect instances of a target behavior when using an operational definition compared to a pinpoint. The operational definition showed an average accuracy of 75% (range = 19–100%), whereas the pinpoint had an average accuracy of 93% (range = 83–100%). Figure 2 presents the results in a column graph (Harris, 1999).

Statistical Significance of Overall Accuracy

When exploring the data, we detected two outliers in a box plot that occurred more than 1.5 box lengths from the edge (Laerd Statistics, 2015). We chose not to remove the data points because the values did not significantly impact the main findings. The difference in scores between the operational definition and pinpoint lacked a normal distribution, as assessed by Shapiro–Wilk's test ($p = 0.030$).

The Wilcoxon signed-rank test sought to determine the effect of behavior definition formats on participant agreement with the "gold standard." Of the 28 participants recruited for the study, the pinpoint elicited an increase in percent agreement for 24 participants compared to the operational definition, whereas four participants experienced no improvement. The Wilcoxon signed-rank test found a statistically significant increase in participant percent agreement (*Mdn* = 15%) when participants used the pinpoint (*Mdn* = 95.0%) compared to the operational definition (*Mdn* = 80.5%), $z = 3.97$, $p < 0.001$.

Individual Participant Accuracy

Figure 3 shows another column graph with the percent agreement versus the gold standard for each participant's response to each target behavior description. Three participants obtained 100% accuracy with operational definitions, while five had 100% accuracy with the pinpoint. Of the eight participants who achieved 100% accuracy, four (i.e., participants 1, 6, 19, and 25) had the other score above 90%. Two other participants, 2 and 9, had 80% or higher on the measure. Therefore, six of

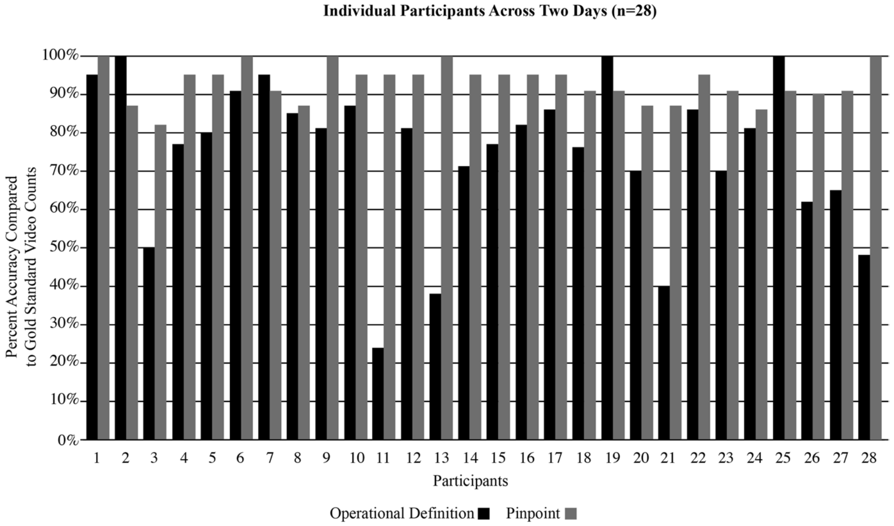

Fig. 3 A column graph showing each individual participants' performance under operational definition and pinpoint conditions

the eight participants could be classified as generally good at behavioral detection. Overall, the data show most participants had lower accuracy with the operational definition compared to the pinpoint.

Statistical Significance of Individual Accuracy

Twenty-four participants (86%) obtained a higher percent agreement when presented with the pinpoint than the operational definition condition. A paired t-test revealed the difference in percent agreements for each group between first period and second period to Group 1 ($M = 0.15$, SD = 0.09) and Group 2 ($M = 0.26$, SD = 0.22) not significant, $M = -0.12$, 95% CI [−0.25, 0.01], $t(26) = -1.84$, $p = 0.078$.

Furthermore, participant's produced more false positives (i.e., identified that a target behavior occurred when one did not occur) when responding using a pinpoint ($M = 0.1$, $SD = 0.04$) than an operational definition ($M = 0.05$, SD = 0.05), a statistically significant mean increase of 5%, 95% CI [−0.032, 0.080], $t(27) = 4.774$, $p < 0.0001$, $d = 0.90$. For false negatives, (i.e., failed to identify a target behavior when a target behavior occurred), higher rates appeared in the operational definition condition ($M = 0.35$, SD = 0.25) and then in the pinpoint condition ($M = 0.11$, SD = 0.09) with a statistically significant mean increase of 24%, 95% CI [0.147, 0.337], $t(27) = 5.228$, $p < 0.0001$, $d = 0.98$. The differences between the two variables appear in Table 2.

Discussion

Operational definitions play a pivotal role in the study behavior, most notably in applied research. Yet concerns surround operationalization that include translation validity, a lack of an agreed method for producing valid and reliable definitions, a dearth of empirical studies substantiating operational definitions as the best means for labeling target behaviors, and inconsistency and variability across definitions for the same target behavior (Alberto et al., 2022; Breitborde et al., 2009; Slife et al., 2016; Spira et al., 2015). A practical issue with operational definitions focuses on detection accuracy (Smith et al., 2013). Pinpoints likewise have limited research but do offer an alternative to operational definitions as they depict the behavior of interest directly and purportedly produce high detection accuracy (Kubina, 2019). The present study marks the first of its kind comparing accuracy detection between operational definitions and pinpoints.

Table 2 Observation error type for operational definitions and pinpoints

	Operational definition (%)	Pinpoint (%)
False-positive rate	5	10
False-negative rate	35	11

The experimental results (Fig. 3) show differences in the overall accuracy of the two different methods for describing the self-injurious target behavior: Operational definitions resulted in less precise detection accuracy than the pinpoints (i.e., 75% vs. 93%). Seventy-five percent accuracy translates into one incorrect observation out of every four observations, an unacceptably high number. Ninety-three percent accuracy, on the other hand, means one incorrect observation out of every ten observations; still not perfect detection accuracy but much closer to 100%. Behavior analysts and behavior support staff rely on accurate monitoring of target behaviors for treatment programs and decision-making could change depending on which method engenders finer behavior detection.

Operationally defined behaviors enjoy the imprimatur of high quality. Mayer et al. (2019) suggest operationally defined behaviors enhance outcome monitoring due to their potential for higher objectivity. However, operational definitions can pose problems from a practical standpoint due to their word count or the number of conditions associated with each definition. The 30-word operational definition for self-injurious behavior in the present study consisted of "using an open palm, closed fist, or object" (Condition 1, the forms and objects representative of the construct translated into a topography) "to make contact" (Condition 2, the action of the construct) "with any area of the head or face" (Condition 3, contextualization of the construct action) "with sufficient force to move the head at least 1 in" (Condition 4, observable criteria substantiating an instance of conditions 2 and 3) (Courtemanche et al., 2018, p. 1100).

Potentially, the burden of memorizing all four conditions in the 30-word operational definition may have affected observational accuracy. Conversely, the 6-word pinpoint "Hits head with hand or fist" had only two conditions; a "movement cycle" or target action and the description of the object receiving the action. The emphasis on a clear beginning and ending, a cycle, and a repeatable nature offer another possible advantage for the pinpoint (Neely, 2019). The combination of an easily recalled descriptor, clear cycles indicating the onset and conclusion of an instance of action, and the identifiable repeatability could, in part, explain the high detection accuracy associated with the pinpoint and the lower performance for operational definitions.

These data show another difference between operational definitions and pinpoints, namely the overall consistency with making observations. The standard deviation and the range shown in Fig. 2 indicate a more considerable variance range for the operational definition than the pinpoint. A benefit of reliability appears in greater treatment integrity, facilitating improved clinical outcomes (Gambrill, 2012; Smith et al., 2013). The pinpoint's previously listed advantages, less memorization, recognizable cycles, and repeatability may have led to higher consistency.

The overall rate for true positives and true negatives (i.e., correct detection of behavior) appears higher for pinpoints than operational definitions (Fig. 2). Yet, the rate for false positives and false negatives occurred differentially among the two variables. False positives manifested twice as much with pinpoints (i.e., 10%) compared to operational definitions (i.e., 5%). False negatives happened three times more with operational definitions (i.e., 35%) when compared to pinpoints (i.e., 11%). In clinical settings, a high preponderance of false negatives masks the severity of target behavior. False negatives for the target behavior may lead a clinician to judge the behavior as less severe than its actual occurrence. Therefore, a clinician might forgo or recommend a weaker course of action due to the faulty assumptions presented by the data (Gambrill, 2012). Still, how behavior analysts react to false positives or negatives, or type 1 and type 2 errors, and different magnitudes of false positives and negatives in a dataset require further analysis.

The present study demonstrated operational definitions performed poorly regarding detection accuracy, and pinpoints did better, though not perfect. Regardless of the precision and care excised in operationalization, the resultant definition will always have some degree of transitional validity. Self-injurious behavior differs from traditional constructs such as anxiety, love, body image, or type A personality (Burke et al., 2012; Frías et al., 2015; Shaw & Dimsdale, 2007). Self-injurious behavior may not represent a true construct but instead a general action label in need of an operational definition to clarify the description of the target behavior. Self-injurious behaviors require translation into measurable and observable components so observers can all detect the same actions (Mayer et al., 2019).

By their very nature, however, general action labels, like all constructs, have operational definitions that vary from behavior analyst to behavior analyst and researcher to researcher. Furthermore, universal or even widespread agreement to definitional conditions such as "… to make contact with any area of the head or face *with sufficient force to … move the head at least 1 in*" (Courtemanche et al., 2018, p. 1100, italics added for emphasis) will appreciably differ and appear at the idiosyncratic dictate of the definition author. For example, "Head hitting was operationally defined as any hand-to-head contact toward the right side of his forehead that was repeated more than three times. This operational definition was chosen to eliminate other responses such as head scratching" (Patel et al., 2000, p. 396–397). The absence of a force condition and the specification of "repeated three times" seems like a very different sort of self-injurious behavior than the other definition.

On the other hand, pinpoints do not rely on constructs or general action labels with widely variant topographies. Pinpoints have a specific framework (i.e., a movement cycle + context for its occurrence) that directly reflect the target behavior. Research has shown that professionals who work in behavior-intensive schools, such as in the present study, can efficiently and effectively learn how to identify movement cycles (Kubina et al., 2016). The data for the current study suggest the addition of context to create pinpoints yield high accuracy. Therefore, behavior analysts may wish to further explore pinpoints in addition to operational definitions for detection accuracy.

Limitations

The present study has several noteworthy limitations. For instance, the operational definition may not have served as the best representation of head hitting. We reviewed multiple articles and found variability in all definitions, including examples and nonexamples (see Table 1), and others without a physical element (i.e., "move the head at least one inch") or representative samples of head hitting. Additionally, we removed the sound detail that appeared in the original definition due to other factors (e.g., poor sound quality of the extracted videos, extraneous sounds that competed with head hits). A different operational definition may have produced different results. Furthermore, the exclusion of the sound could affect external validity.

Additionally, as part of school's onboarding process and professional development opportunities, the participants did receive training in behavior analysis and data collection (i.e., detecting and observing behavior). The time spent learning the operational definition and pinpoint appeared similar to previous study which also examined behavioral definitions (i.e., Smith et al., 2013). However, more time familiarizing and practicing each definition could have affected their performance. We could not find examples in textbooks or published articles that recommend an optimal time for learning operational definitions or pinpoints.

Future Directions

The present study revealed pinpoints provided higher detection accuracy than operational definitions. Additional studies must examine if other general action labels and operationally defined constructs perform similarly to their pinpoint alternatives. Many differences exist in how behavior analysts and researchers create operational definitions due to a lack of rules or a standardized format. Perhaps some other form of operational definitions would perform differently than the description of self-injurious behavior used in the present study? The different types of options used in creating operational definitions by Alberto et al., (2022) demonstrate multiple form factors. The present study used one form and a form that generated fewer total words than the operational definition. Therefore, readers should exercise caution weighing the relative value of operational definitions. Moreover, this study marks the first empirical evaluation of the detection accuracy of a pinpoint. More research will bear the value of a pinpoint's detection accuracy and clinical usefulness.

Other questions surround the practical differences between operational definitions and pinpoints. For instance, a common practice involves counting multiple behaviors under one label. Inappropriate mealtime behavior, as an example, had several behaviors in its definition and appeared "each time the food, drink, or utensil was within the child's reach, and the child's (a) mouth turned 45 degrees or moved 5 cm in any direction except toward the utensil; (b) hand, arm, or anything in their hand touched the food, drink, utensil, or feeder's arm; or (c) hand, arm, or anything in their hand (except the utensil) contacted their lips" (Andersen et al., 2022, p. 268). Researchers and practitioners have distinctive reasons for selecting multiple behaviors; the behaviors could form a response class, appear due to other researchers establishing such behaviors as the de facto standard, or perhaps serve as the most pressing behaviors for the specific research participant.

Pinpoints could also represent multiple behaviors but would do so in a different fashion. "Volleys tennis ball during a match" illustrates a pinpoint that would involve using a specific grip (e.g., Continental grip), running to where the ball bounces, and swinging a tennis racket—all three behaviors form one cycle with an explicit beginning and ending. More specificity could also occur with the pinpoint if someone wanted to count a specific swing during a volley (e.g., forehand versus backhand). To target inappropriate mealtime behavior as previously described (Andersen et al., 2022), the researcher or practitioner must create three separate pinpoints. Future research could examine the need to target and address behaviors that occur in a response class and the detection accuracy for large sets of behavior in an operational definition versus fewer targets in one to two pinpoints or pinpoints representing more than one discrete action.

Future research could also explore how a pinpoint contributes to the function of behavior compared to an operational definition. Behavior analysis has a rich area of practice ranging from home and school settings to business and other professional organizations. The variety of operational definitions and pinpoints could lead to future intriguing and pivotal studies. And last, the present study systematically replicated previous research (i.e., Smith et al., 2013) that used one definition versus a second definition to ascertain differences in detection accuracy. Future studies that present multiple operational definitions compared to multiple pinpoints would show further consistency in participants' ability to accurately detect the target behavior and error rates.

Declarations

Conflict of interest None of the authors have any conflicts of interest to disclose.

Ethical Approval All procedures performed in studies involving human participants met the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments.

Consent to participants We obtained informed consent from all participants before including them in the study.

References

Alberto, P. A., Troutman, A. C., & Axe, J. (2022). *Applied behavior analysis for teachers* (10th ed.). Pearson Education.

Andersen, A. S., Hansen, B. A., & Peterson, K. M. (2022). An evaluation of trial-based functional analyses of inappropriate mealtime behavior. *Journal of Applied Behavior Analysis, 55*, 264–289. https://doi.org/10.1002/jaba.888

Austin, J. L., & Soeda, J. M. (2008). fixed-time teacher attention to decrease off-task behaviors of typically developing third graders. *Journal of Applied Behavior Analysis, 41*(2), 279–283. https://doi.org/10.1901/jaba.2008.41-279

Beserra, V., Nussbaum, M., & Oteo, M. (2019). On-task and off-task behavior in the classroom: A study on mathematics learning with educational video games. *Journal of Educational Computing Research, 56*(8), 1361–1383. https://doi.org/10.1177/0735633117744346

Bird, F., Wachtel, L. E., Henry, M., Gold, J., Fernandez-Robles, C., Orchanian, S., Shlesinger, A., & Luiselli, J. K. (2020). Treatment and maintenance effects of behavioral intervention and electroconvulsive therapy (ECT) in a man with catatonia, life-threatening self-injury, and autism spectrum disorder. *Advances in Neurodevelopmental Disorders, 5*(2), 135–143. https://doi.org/10.1007/s41252-020-00189-0

Boring, E. G. (1945). The use of operational definitions in science. *Psychological Review, 52*(5), 243–245. https://doi.org/10.1037/h0054934

Breitborde, N. J. K., Srihari, V. H., & Woods, S. W. (2009). Review of the operational definition for first episode psychosis. *Early Intervention in Psychiatry, 3*, 259–265. https://doi.org/10.1111/j.1751-7893.2009.00148.x

Bridgman, P. W. (1945). Some general principles of operational analysis. *Psychological Review, 52*(5), 246–249. https://doi.org/10.1037/h0060381

Burke, N. L., Schaefer, L. M., & Thompson, J. K. (2012). *Body image* (2nd ed., pp. 365–371). Elsevier Inc. https://doi.org/10.1016/B978-0-12-375000-6.00066-5

Carr, J. E., & Nosik, M. R. (2017). Professional credentialing of practicing behavior analysts. *Policy Insights from the Behavioral and Brain Sciences, 4*(1), 3–8. https://doi.org/10.1177/2372732216685861

Cooper, J. O., Heron, T. E., & Heward, W. L. (2020). *Applied behavior analysis* (3rd ed.). Pearson Education.

Courtemanche, A. B., Lloyd, B. P., & Tapp, J. T. (2018). A descriptive analysis of self-injury in community settings: Exploring behaviour-behaviour contingencies. *Journal of Intellectual Disability Research, 62*(12), 1097–1107. https://doi.org/10.1111/jir.12485

Ehrenfeld, N., & Horn, I. S. (2020). Initiation-entry-focus-exit and participation: A framework for understanding teacher groupwork monitoring routines. *Educational Studies in Mathematics, 103*(3), 251–272. https://doi.org/10.1007/s10649-020-09939-2

Feigl, H. (1945). Operationism and scientific method. *Psychological Review, 52*(5), 250–259. https://doi.org/10.1037/h0056755

Frías, M. T., Shaver, P. R., & Mikulincer, M. (2015). Chapter 15—Measures of adult attachment and related constructs. In G. Boyle, D. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 417–447). Elsevier.

Gambrill, E. (2012). *Critical thinking in clinical practice: Improving the quality of judgements and decisions* (3rd ed.). Wiley.

Gerow, S., Radhakrishnan, S., Davis, T. N., Zambrano, J., Avery, S., Cosottile, D. W., & Exline, E. (2021). Parent-implemented brief functional analysis and treatment with coaching via telehealth. *Journal of Applied Behavior Analysis, 54*(1), 54–69. https://doi.org/10.1002/jaba.801

Godwin, K. E., Almeda, M. V., Seltman, H., Kai, S., Skerbetz, M. D., Baker, R. S., & Fisher, A. V. (2016). Off-task behavior in elementary school children. *Learning and Instruction, 44*, 128–143. https://doi.org/10.1016/j.learninstruc.2016.04.003

Gregori, E., Rispoli, M., Gerow, S., Gerow, S., & Lory, C. (2018). Treatment of self-injurious behavior in adults with intellectual and developmental disabilities: A systematic review. *Journal of Developmental and Physical Disabilities, 30*(1), 111–139. https://doi.org/10.1007/s10882-017-9568-7

Hagopian, L. P., Rooker, G. W., & Zarcone, J. R. (2015). Delineating subtypes of self-injurious behavior maintained by automatic reinforcement. *Journal of Applied Behavior Analysis, 48*(3), 523–543. https://doi.org/10.1002/jaba.236

Harris, R. L. (1999). *Information graphics: A comprehensive illustrated reference*. Oxford University Press.

Huisman, S., Mulder, P., Kuijk, J., Kerstholt, M., van Eeghen, A., Leenders, A., van Balkom, I., Oliver, C., Piening, S., & Hennekam, R. (2018). Self-injurious behavior. *Neuroscience and Biobehavioral Reviews, 84*, 483–491. https://doi.org/10.1016/j.neubiorev.2017.02.027

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities, 2*(1), 3–20. https://doi.org/10.1016/0270-4684(82)90003-9

Johnston, J. M., Pennypacker, H. S., & Green, G. (2020). *Strategies and tactics of behavioral research and practice* (4th ed.). Routledge.

Kazdin, A. E. (2020). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kraemer, E. E., Davies, S. C., Arndt, K. J., & Hunley, S. (2012). A comparison of the mystery motivator and the get 'em on task interventions for off-task behaviors. *Psychology in the Schools, 49*(2), 163–175. https://doi.org/10.1002/pits.20627

Krathwohl, D. R. (2009). *Methods of educational and social science research: The logic of methods* (3rd ed.). Waveland Press Inc.

Kubina, R. M., & Yurich, K. K. L (2012). *The precision teaching book*. Greatness Achieved.

Kubina, R. M. (2019). *The precision teaching implementation manual*. Greatness Achieved.

Kubina, R. M., Yurich, K. L., Durica, K. C., & Healy, N. M. (2016). Developing behavioral fluency with movement cycles using SAFMEDS. *Journal of Behavioral Education, 25*, 120–141. https://doi.org/10.1007/s10864-015-9232-1

Laerd Statistics (2015). *Wilcoxon signed-rank test using SPSS Statistics. Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.). Routledge.

Lindsley, O. R. (1972). From Skinner to precision teaching: The child knows best. In J. B. Jordan & L. S. Robbins (Eds.), *Let's try doing something else kind of thing* (pp. 1–12). Arlington, VA: Council for Exceptional Children.

Lindsley, O. R. (1990). B.F. Skinner (1904-1990) Thank you Grandpa Fred. *Journal of Precision Teaching, 8*(1), 5–11.

Lindsley, O. R. (1991). Precision teaching's unique legacy from B. F. Skinner. *Journal of Behavioral Education, 1*(2), 253–266. https://doi.org/10.1007/BF00957007

Linscheid, T. R., Iwata, B. A., Ricketts, R. W., Williams, D. E., & Griffin, J. C. (1990). Clinical evaluation of the self-injurious behavior inhibiting system (SIBIS). *Journal of Applied Behavior Analysis, 23*(1), 53–78. https://doi.org/10.1901/jaba.1990.23-53

Linscheid, T. R., Pejeau, C., Cohen, S., & Footo-Lenz, M. (1994). Positive side effects in the treatment of SIB using the self-injurious behavior inhibiting system (SIBIS): Implications for operant and biochemical explanations of SIB. *Research in Developmental Disabilities, 15*(1), 81–90. https://doi.org/10.1016/0891-4222(94)90040-X

Mayer, G. R., Sulzer-Azaroff, B., & Wallace, M. (2019). *Behavior analysis for lasting change* (4th ed.). Sloan.

Menin, D., Guarini, A., Mameli, C., Skrzypiec, G., & Brighi, A. (2021). Was that (cyber) bullying? Investigating the operational definitions of bullying and cyberbullying from adolescents' perspective. *International Journal of Clinical and Health Psychology, 21*(2), 1–8. https://doi.org/10.1016/j.ijchp.2021.100221

Neely, M. D. (2019). Precision teaching tools: A. The plan book. In N. Haring, M. White, & M. Neely (Eds.), *Precision teaching—A practical science of education* (pp. 44–58). Sloan.

Novack, M. N., & Dixon, D. R. (2019). Predictors of burnout, job satisfaction, and turnover in behavior technicians working with individuals with autism spectrum disorder. *Review Journal of Autism and Developmental Disorders, 6*(4), 413–421. https://doi.org/10.1007/s40489-019-00171-0

O'Rourke, S., Richling, S., Brogan, K., McDougale, C., & Rapp, J. T. (2019). Tolerance training with adolescents in a residential juvenile facility. *Behavior Modification*. https://doi.org/10.1177/0145445519890261

Patel, M. R., Carr, J. E., Kim, C., Robles, A., & Eastridge, D. (2000). Functional analysis of aberrant behavior maintained by automatic reinforcement: Assessments of specific sensory reinforcers. *Research in Developmental Disabilities, 2*(5), 393–407. https://doi.org/10.1016/S0891-4222(00)00051-2

Piantadosi, S. (2005). Crossover designs. In S. Piantadosi (Ed.), *Clinical trials: A methodologic perspective* (2nd ed., pp. 515–528). Wiley.

Putt, M. E., & Chinchilli, V. M. (2004). Nonparametric approaches to the analysis of crossover studies. *Statistical Science, 19*(4), 712–719. https://doi.org/10.1214/088342304000000611

Reed, J. F. (2004). Analysis of two-treatment, two-period crossover trials in emergency medicine. *Annals of Emergency Medicine, 43*(1), 54–58. https://doi.org/10.1016/S0196-0644(03)00661-9

Robinson, J., Gershwin, T., & London, D. (2019). Maintaining safety and facilitating inclusion: Using applied behavior analysis to address self-injurious behaviors within general education classrooms. *Beyond Behavior, 28*(3), 154–167. https://doi.org/10.1177/1074295619870473

Rooker, G. W., Hausman, N. L., Fisher, A. B., Gregory, M. K., Lawell, J. L., & Hagopian, L. P. (2018). Classification of injuries observed in functional classes of self-injurious behaviour. *Journal of Intellectual Disability Research, 62*(12), 1086–1096. https://doi.org/10.1111/jir.12535

Shaw, W. S., & Dimsdale, J. E. (2007). Type A personality, type B personality. In G. Fink (Ed.), *Encyclopedia of Stress* (2nd ed., pp. 782–786). Academic Press.

Shore, B. A., Iwata, B. A., Lerman, D. C., & Shirley, M. J. (1994). assessing and programming generalized behavioral reduction across multiple stimulus parameters. *Journal of Applied Behavior Analysis, 27*(2), 371–384. https://doi.org/10.1901/jaba.1994.27-371

Sibbald, B., & Roberts, C. (1998). Understanding controlled trials crossover trials. *BMJ, 316*(7146), 1719–1719. https://doi.org/10.1136/bmj.316.7146.1719xw

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Copley.

Skinner, B. F. (1945). The operational analysis of psychological terms. *Psychological Review, 52*, 270–277. https://doi.org/10.1037/h0062535

Slife, B. D., Wright, C. D., & Yanchar, S. C. (2016). Using operational definitions in research: A best-practices approach. *The Journal of Mind and Behavior, 37*(2), 119–140.

Smith, G. D., Lambert, J. V., & Moore, Z. (2013). Behavior description effect on accuracy and reliability. *The Journal of General Psychology, 140*(4), 269–281. https://doi.org/10.1080/00221309.2013.818525

Smith, R. G., Lerman, D. C., & Iwata, B. A. (1996). Self-restraint as positive reinforcement for self-injurious behavior. *Journal of Applied Behavior Analysis, 29*(1), 99–102. https://doi.org/10.1901/jaba.1996.29-99

Spira, D., Buchmann, N., Nikolov, J., Demuth, I., Steinhagen-Thiessen, E., Eckardt, R., & Norman, K. (2015). Association of low lean mass with frailty and physical performance: A comparison between two operational definitions of sarcopenia-Data from the Berlin Aging Study II (BASE-II). *The Journals of Gerontology: Series A, 70*(6), 779–784. https://doi.org/10.1093/gerona/glu246

Szymanski, L., Kedesdy, J., Sulkes, S., Cutler, A., & Stevens-Our, P. (1987). Naltrexone in treatment of self injurious behavior: A clinical study. *Research in Developmental Disabilities, 8*(2), 179–190. https://doi.org/10.1016/0891-4222(87)90002-3

Tate, B. G., & Baroff, G. S. (1966). Aversive control of self-injurious behavior in a psychotic boy. *Behaviour Research and Therapy*, *4*(4), 281–287. https://doi.org/10.1016/0005-7967(66)90024-6

Trikalinos, T. A., Balion, C. M., Coleman, C. I., Griffith, L., Santaguida, P. L., Vandermeer, B., & Fu, R. (2012). Chapter 8: Meta-analysis of test performance when there is a “Gold standard”. *Journal of General Internal Medicine*, *27*(S1), 56–66. https://doi.org/10.1007/s11606-012-2029-1

Vosaic, (2020). *Vosaic* [Computer software]. https://vosaic.com/

White, O. R., & Haring, N. G. (1980). *Exceptional Teaching* (2nd ed.). Merrill.

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