Practical Implications of Data Reliability and Treatment Integrity Monitoring

Timothy R. Vollmer, Ph.D., BCBA and Kimberly N. Sloman, Ph.D., BCBA, University of Florida
Claire St. Peter Pipkin, Ph.D., BCBA , West Virginia University

ABSTRACT

Data reliability and treatment integrity have important implications for clinical practice because they can affect clinicians' abilities to accurately judge the efficacy of behavioral interventions. Reliability and integrity data also allow clinicians to provide feedback to caregivers and to adjust interventions as needed. We present reasons why reliability and integrity measures are paramount in clinical work, discuss events that may result in decreased reliability or integrity, and provide several efficient means for collecting data and calculating reliability and integrity measures.
Descriptors: Data analysis, integrity, reliability

It is standard practice to record *data reliability* (i.e., interobserver agreement) when conducting applied behavioral experiments (Hartmann, 1977). It is not standard practice to record *treatment integrity* in applied behavioral experiments, but there have been strong calls to do so, along with some recent evidence that the practice is increasing in frequency (McIntyre, Gresham, DiGennaro, & Reid, 2007). However, there has been little discussion about the importance of these measures in everyday practice of behavior analysis. The purpose of this paper is to provide a brief background on types of reliability and integrity measures, a rationale for the use of these measures in clinical settings, and some possible methods to collect reliability and integrity data.

Various kinds of reliability measures can be taken, but in this paper we are referring specifically to the extent to which two observers agree on the occurrence or nonoccurrence of events. For example, if person A records an instance of aggression between 2:30 p.m. and 2:35 p.m., does person B also record an instance of aggression during that time frame? Do the observers agree that the episode did or did not occur? By treatment integrity, we mean the extent to which behavioral procedures are conducted according to a behavior change plan (Gresham, Gansle, Noell, Cohen, & Rosenblum, 1993). For example, if the behavior plan states that a reinforcer should be delivered after some specified instance of vocal communication, is the reinforcer actually delivered?

In the course of our behavior analytic practice, we frequently evaluate the implementation of behavioral procedures in service settings, schools, homes, and other settings. Often, when we begin to detail the process of collecting reliability data and treatment integrity data, we hear something akin to the following complaint: "We are not conducting research here. I know you are researchers and for you that kind of thing is important, but we are running a treatment center, not conducting an experiment." Such comments have come from a range of people including teachers, behavior analysts, sophisticated parents, and others. In other words, many skilled people are conducting practice without data reliability and treatment integrity monitoring. We view this as a potentially dangerous practice.

A failure to collect data reliability and treatment integrity measures is potentially dangerous because life-changing decisions are made based on the assumption that the data reported are reasonably accurate and based on the assumption that the prescribed procedures were conducted as specified. Some life-changing decisions that arise from these assumptions include residential placement, the use of restrictive behavioral procedures, changes or lack thereof in psychotropic medication, use of restrictive or labor intensive staffing, and so on. It seems clear that few would question the appropriateness of data reliability and treatment integrity if the problem was medical rather than behavioral. Consider two medical analogies:

- Patient A has severe seizures and is therefore prescribed medication Z as treatment. Patient A's parents are asked to record all instances of seizures before the introduction and after the introduction of medication Z. Suppose the parents are reasonably diligent and accurately recording seizures prior to the medication, but slack off a bit and forget to record many of the seizures after the introduction of medication Z. At the next medical appointment, based on the parents' data, Patient A's medical team concludes that medication Z was effective and the patient shall remain on the medication. In truth, data reliability checks would have shown that the recording of seizures had slacked off and there was no real change in frequency. Medication Z was ineffective but the data suggested otherwise. Patient A receives an ineffective medication and a potentially effective medication (say, medication Y) is left on the shelf.
- Patient B has severe seizures and is therefore prescribed medication X as treatment. Patient B's nurse is asked to administer the medication X twice

daily. Suppose the nurse frequently forgets to give the medication but her data records of seizure episode frequency are reasonably accurate (showing no change, because the medication is given no chance to work). At the next medical appointment, Patient B's medical team concludes that Medication X was ineffective and is moved to prescribe Medication W as an alternative, and Medication W is known to have serious side effects. In truth, the medication may have been effective if administered as prescribed and now Patient B is receiving a more dangerous medication.

Such examples are relatively straightforward because it is not difficult to understand the need for (a) accurately monitoring a medical condition that is being treated via medication, and (b) accurately administering medication. It is simple enough to insert behavior and behavioral procedures into parallel examples, as follows:

- Person C displays severe self-injury and therefore receives a thorough behavioral assessment by a qualified team. A procedure based on differential reinforcement is prescribed as a result of the assessment outcome. Person C's parents are given instructions to conduct the procedure and to record data on self-injury prior to and following the implementation of treatment. The parents are diligent and reasonably accurate with data collection prior to treatment, but they slack off a bit following the initiation of treatment and forget to record many instance of self-injury. At the next interdisciplinary professional team meeting, the team concludes that the behavioral treatment was effective based on the parents' data, and no changes are made. In truth, data reliability checks would have shown that the behavioral treatment was ineffective. The differential reinforcement procedure was not working and other potentially effective behavioral procedures (such as other variations on differential reinforcement) are left on the "shelf."
- Person D displays severe self-injury and therefore receives a thorough behavioral assessment by a qualified team. A procedure based on differential reinforcement is prescribed as a result of the assessment outcome. Person D's teacher is asked to implement the procedure but frequently forgets to do so, and in the process reinforces self-injury and places alternative behavior on extinction. At the next interdisciplinary professional team meeting, the team concludes that the behavioral treatment was ineffective and they prescribe a potentially dangerous psychotropic medication, contingent physical restraint, and extra staffing. In truth, treatment integrity checks would have shown that the procedure was not implemented correctly and it may well have been effective if conducted with good integrity. The person now receives dangerous, intrusive, and costly interventions.

In these examples, data reliability errors resulted in "false positive" treatment outcomes (falsely showing a good treatment effect) and treatment integrity errors resulted in "false negative" outcomes (falsely showing no treatment effect). These examples were meant to highlight some of the implications of data reliability and treatment integrity monitoring. Our general thesis is that measuring reliability and integrity is inherently important. In addition, there are several advantages to such an approach that may have practical utility on a day-to-day basis. Next, we will present some practical usages of data reliability and treatment integrity monitoring.

Practical Usage

One practical usage of data reliability and treatment integrity monitoring is to provide immediate feedback to the data collector and implementer of the procedure. The feedback should take two forms: (a) positive feedback for correct data recording and/or procedural implementation, and (b) corrective feedback for incorrect data recording and/or procedural implementation (DiGennaro, Martens, & Kleinmann, 2007; Sulzer-Azaroff & Mayer, 1991). Of course the incorrect data or procedural implementation may not be the "fault" of the data collector or treatment implementer, such as when there are poor behavioral definitions. In such cases, the data collector/procedure implementer should not receive positive or corrective feedback but should be invited to help review definitions and other sources of error. When feedback is given, we recommend that any opportunity for positive feedback should be seized upon. For example, the person monitoring the data should avoid statements such as "Well, that was a waste of time, the behavior did not even occur so we could not compare our data." Rather, if both observers did not record an instance of behavior, the monitor can say, "Great, we both recorded that the behavior did not occur. That goes down as an agreement and we were successful today." Similarly, there is almost always some client/student appropriate behavior to reinforce, so the presence or absence of the targeted problem behavior should not be the only recorded behavior for which the monitor provides feedback.

We have found that if the corrective feedback occurs with great frequency in relation to the positive feedback, the monitor can become a conditioned aversive stimulus. That is, data collectors and treatment implementers may begin to escape or avoid monitoring sessions. On the other hand, when the monitor frequently points out correct data recording and procedural implementation, the sessions should be favorable for the primary data collector/procedure implementer. It may be important to schedule observations during periods when the behavior is most likely to occur, in order to provide more opportunities for comparison and feedback. For example, if target behavior is maintained by escape from instructions, the observation should be scheduled for instructional sessions. Strategic scheduling of observations may be especially important for low rate behavior in order to increase observation opportunities.

A second practical usage is to provide delayed and cumulative performance feedback to data collectors/procedure implementers (Noell et al., 2000). This function is similar to the immediate feedback discussed above, but relies on the added feature of long-term performance trends. With the same caveats as discussed above (such as other reasons for poor reliability and integrity including poorly phrased definitions), delayed feedback could take two general forms: (a) positive feedback in the form of recognition, promotion, and praise, or (b) corrective feedback in the form of additional training or further detailing of procedures or supervisor meetings (Noell et al.). Some excellent uses of delayed positive feedback for cumulative performance include public recognition at a staff or parent meeting (e.g., “Mrs. Smith has been providing care for a child with very dangerous behavior; I am happy to report that her data reliability scores have exceeded 90% for the past three months and her treatment implementation scores were 100% for last month!”); public recognition via awards; acknowledgement on a website or in a newsletter or newspaper; and so on, including private recognition in a written or oral performance evaluation.

A third practical usage relates to clinical decision-making. Changes in behavioral procedures should be informed by reliability and treatment integrity data, as exemplified by the hypothetical cases presented earlier. For example, if there is an increase in problem behavior rates simultaneous with improved data reliability scores, it is possible that data collectors are simply getting better at data collection and, therefore, increased problem behavior rates may not present a need for changed procedures. In the case of treatment integrity, it is possible that poor treatment effects are not due to a poor treatment per se, but due to a treatment that is not being implemented sufficiently. Figure 1 shows a hypothetical example of using integrity measures to determine a need for booster training (for an actual example, see Vollmer, Marcus, & LeBlanc, 1994). Thus, a behavior analyst should be equipped with both reliability and treatment integrity data whenever critical clinical decisions are being made.

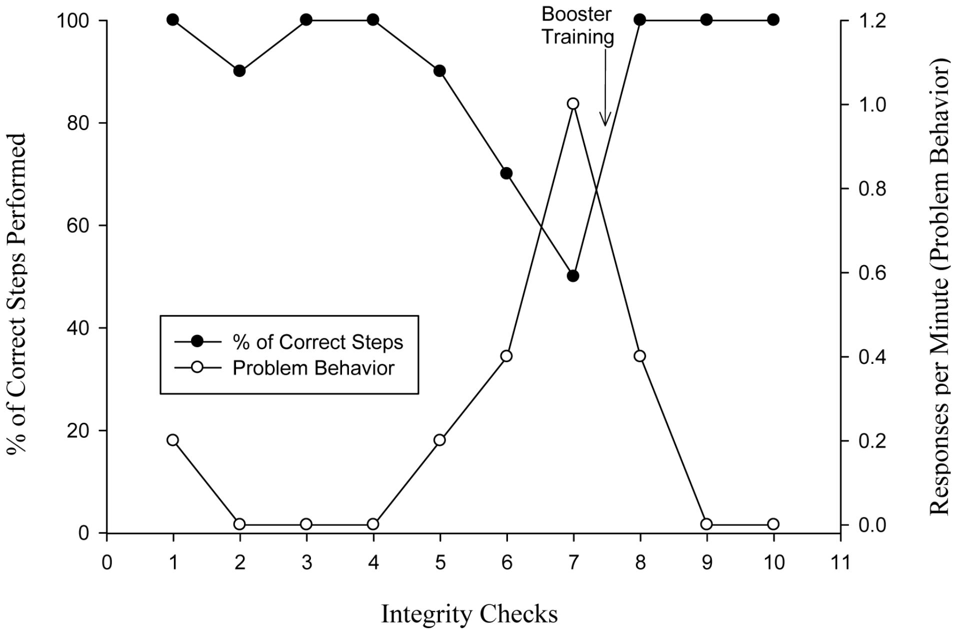


Figure 1. Hypothetical data showing the interaction between the percentage of correct steps completed (treatment integrity, shown in the filled circles) and child problem behavior (shown in the open circles). Child problem behavior increases as treatment integrity decreases; a booster training (shown by the arrow) results in increased treatment integrity and regained treatment effects.

If reliability and integrity measures are solid, then good clinical decisions can be made based on a proper evaluation of treatment effects or lack thereof.

Some Caveats about Reliability and Procedural Integrity

It is important to note that a high reliability score does not necessarily equate to high accuracy. Clearly, two observers could be wrong about the same thing (Hawkins & Dotson, 1975). Also, because some reliability measures tend to be either more conservative or more liberal than others, there is no “magic” score that would indicate good reliability. Because of these caveats related to data reliability, we recommend using the measures to indicate when something is clearly wrong. In other words, one should not necessarily be comforted by a high percentage of agreement, but one should certainly be concerned by a low percentage of agreement.

An important caveat about treatment integrity is that different procedures require different levels of correct implementation. For example, an occasional error on an extinction procedure equates to an intermittent schedule of reinforcement. Suppose a parent correctly implements extinction during 95% of episodes of child night time disruptive behavior. This means that the behavior is reinforced on a variable ratio (VR) 20 schedule, which could maintain the problem behavior. Thus, an integrity score that looks and sounds “high” may be very bad, depending on the procedure. Alternatively, some procedures may not require high levels of integrity to be successful. For example, an occasional error on a differential reinforcement of alternative (DRA) behavior schedule might not be damaging if the alternative (desirable) behavior receives more reinforcement than the problem behavior. Suppose a parent reinforces tantrums on a VR 4 schedule (75% integrity if the prescribed intervention is no reinforcement following tantrums) but reinforces appropriate requests for attention on a VR 2 schedule (50% integrity if the prescribed intervention is reinforcement following all appropriate requests for attention). Because the schedule is much richer for appropriate behavior, we can predict based on decades of research on choice behavior that the child would allocate almost all behavior in the direction of appropriate behavior. Thus, what might look and sound like "low" integrity may be very good, depending on the procedure. Our examples using extinction and differential reinforcement are intended to be illustrative rather than comprehensive, as of course behavioral procedures have many complexities that can be relatively sensitive or insensitive to integrity problems (such as different prompting methods, and so on).

A rule of thumb might be to conclude that data reliability and treatment integrity scores should be considered carefully in a context from which these data are collected. How conservative or liberal is the reliability measure? How important is it to record every instance of behavior? What treatment procedure is being used? What is the likely effect of a treatment integrity error given the procedure used?

With one or two exceptions, we have written so far on the assumption that a reliability or integrity error is committed by the primary observer/implementer. This may be true, but it may not be the "fault" of the observer/implementer per se. In the sections that follows we will discuss some common types of errors and then some common reasons for (or sources of) those errors.

Common Reliability and Integrity Errors

There are several possible errors that may contribute to low reliability or integrity scores. The two most basic reliability and integrity errors may be described as errors of omission and commission. Errors of omission occur when observers or personnel implementing behavioral programs do not provide the appropriate response when a specific event occurs. For data reliability, errors of omission may include failing to document a response or environmental event. For treatment integrity, errors of omission may include failing to deliver a reinforcer for an appropriate alternative response in a DRA procedure.

Errors of commission occur when observers or personnel implementing behavioral programs provide a response at an inappropriate time. For data reliability, errors of commission may include recording an event when it did not occur, or recording one event in place of a different event. For example, an observer may record that a child engaged in self-injury when instead he engaged in aggression. For treatment integrity, errors of commission may include delivering some antecedent or consequence at an inappropriate time. For example, a therapist may accidentally deliver a reinforcer after problem behavior in a DRA treatment session.

Some reliability and integrity errors may be subtler than those describe above. For example, two observers may record the same response but at slightly different times. To illustrate, suppose two observers are recording instances of self-injury and reliability is assessed on a minute-by-minute basis. If observer A records an instance of self-injury at the end of minute 5 and observer B records an instance of self-injury at the beginning of minute 6, there would be a lack of agreement within those respective intervals. If this discrepancy occurs frequently throughout the data collection, these errors could result in both poor reliability scores and dissimilar data outcomes. Similarly, integrity errors can be said to occur any time there are discrepancies between the prescribed protocol and the actual implementation of events (Peterson et al., 1982). That is, integrity errors may include inappropriate reinforcer delivery as well as slight changes in the protocol. For example, the errors may include delivery of reinforcers after a delay and the presentation of social cues such as nods or smiles from the therapist.

Sources of Reliability and Integrity Errors

Several possible causes for reliability and integrity errors have been outlined in the literature (e.g. Allen & Warzak, 2000; Kazdin, 1977; Peterson et al.). One main factor influencing these errors may simply be inadequate or incomplete training of the protocols. More specifically, observers may not know how to fill out the data collection forms or use data collection devices, and may also commit errors because they are not aware of the correct definitions of behavior and environmental events. Likewise, persons implementing behavioral programs may not have sufficient information to conduct the protocol.

A second factor influencing integrity and reliability is the complexity of the protocol. For example, if a protocol requires an observer to collect data on numerous responses and environmental events, he or she may be more likely to commit reliability errors. Similarly, integrity errors may be more likely in a case where the person implementing the program has to complete several different steps (e.g. a detailed prompting sequence) across a variety of responses (e.g. both appropriate and inappropriate behavior) or with numerous clients or students. Thus, it is important that both observers and therapists are given clear, detailed, and manageable instructions on the protocol and behavioral definitions. In addition, individuals should be provided with ample time to practice performing the required tasks, and should be provided immediate positive and corrective feedback about their performance during training procedures (i.e., competency based training).

A third factor is the failure to generalize from the training setting. Namely, individuals may be able to perform the skills (data collection or treatment implementation) accurately in the training sessions, but fail to do so in the actual environment. Generalization of the skills may be facilitated by training several different exemplars (e.g. instances of the behavior) and conducting training in several different environments (Stokes & Baer, 1977).

A fourth possible factor influencing reliability and integrity errors has been referred to as a "drift" in performance (e.g., Kazdin, 1977). That is, individuals initially perform the skills as prescribed but then drift or alter their behavior from the original protocol. Careful monitoring of observers and those individuals implementing programs combined with periodic booster training sessions may help to prevent drift from occurring.

A fifth possible factor influencing reliability and integrity errors may be competing environmental contingencies. More specifically, there may be reinforcers for departures from the protocol, punishers in place for adherence with the protocol, or both. For example, a study by O'Leary, Kent, and Kanowitz (1975) showed that observers who had received specific information about the session (e.g. behavior should decrease in the treatment phase) and feedback (e.g. praise for scoring low rates of behavior and reprimands for scoring higher rates of behavior) were biased in their data collection. Likewise, inaccurate reports of low rates of problem behavior by caregivers may be accidentally reinforced by praise and encouragement from a behavior analyst, especially if the behavior analyst is not present when the data collection is taking place. Conversely, reports might be more accurate only when the caregiver is aware that a behavior analyst was currently collecting reliability data (e.g., Brackett, Reid, & Green, 2007). Thus, it would be important in these circumstances to emphasize and praise the accuracy of data collection and refrain from mentioning specific changes in behavior.

Integrity errors may also occur due to competing schedules of reinforcement. For example, a behavior analyst may recommend that parental attention be delivered for appropriate behavior, and not for tantrums. However, the parent may be in a setting (e.g., a grocery store) in which adherence to the program is not reinforced and may actually be punished (e.g., other shoppers giving dirty looks). Therefore, the parent's delivery of attention for tantrums is negatively reinforced and future integrity errors become more likely. Emphasizing accuracy, providing consistent feedback about integrity level, and providing reinforcement for high levels of integrity may be necessary to maintain high levels of integrity (DiGennaro, Martens, & Kleinmann, 2007).

Some Suggestions for Reliability Measures

As mentioned earlier, different methods of calculating reliability may yield more conservative or liberal estimates. In addition, measures vary in their ease of calculation. Thus, practitioners may choose measures based on whether stringent criteria or ease of calculation are desirable, as well as on the type of data that are available.

Reliability measures vary in at least two ways: the size of the time window and the type of data. Larger time windows may make calculations easier than smaller ones. One of the simplest ways of calculating reliability is to count the total number of responses scored (or the total number of intervals containing responses, depending on the data collection system) by each observer throughout the observation period, to divide the smaller number by the larger number, and to then multiply by 100. This yields an overall percentage of agreement for that observation. Whole-session measures are simple to understand and to calculate, but they provide only a liberal estimate of the reliability of the data collection. For an extreme example, one observer could score 10 instances of the target response, then become distracted or fall asleep. The second observer may miss those initial 10 responses, but later record 10 other responses (while the first observer sleeps). A whole-session reliability measure for these two data sets would be 100% because both observers scored 10 responses, but those responses would have occurred at entirely different times.

Using shorter intervals within a longer observation period makes reliability calculations more stringent and improves the confidence that both observers were recording the same instance of behavior. The use of shorter, within-session intervals is sometimes called the proportional method. To calculate proportional agreement, the total observation time is broken into discrete units (intervals). For instance, a 10-min observation might be broken into 60, 10-s intervals. The records for the two observers are then compared within each 10-s interval. For example, if one observer recorded two instances of behavior in the first 10-s interval and a second observer recorded three instances of behavior in the first interval, the reliability for that interval would be 66.7% (two instances divided by three instances and multiplied by 100). Once reliability has been calculated for all intervals in the observation, the scores are averaged to obtain the mean reliability for the entire observation. Although 10-s intervals are common in research, larger intervals such as 1-min or 5-min may be more practical in everyday application.

Proportional reliability has several possible advantages over whole-session reliability. First, proportional measures are more stringent than whole-session measures. By breaking the session into smaller units, interval-by-interval calculations reduce the likelihood of obtaining good reliability when two observers record entirely different responses (as in the example given for whole-session reliability above).

Another method is the exact agreement method, for which the observational intervals are scored as an "agreement" if both observers counted exactly the same number of behavior instances. If they do not agree exactly, the interval is scored as a "disagreement." The number of agreements are then divided by the total number of intervals and converted to a percentage. This method is even more conservative than the proportional method, but it can sometimes be overly conservative. For example, when the observers are just slightly off in their timing, behavior scored in one interval for one observer and in another interval for a second observer produces two disagreement intervals even though both observers were scoring the same behavioral event.

Another method for reliability is used when partial interval or whole interval recording is in place. Partial interval refers to scoring the interval if the behavior occurs at any point in that interval. Whole interval recording refers to scoring the interval if the behavior occurs throughout the interval. Thus, there is no "count" of behavior; the interval is simply scored as "occurrence" or "nonoccurrence." In the case of interval recording, reliability can be calculated by denoting each interval as either an agreement (both observers recorded behavior or did not record behavior) or a disagreement (one observer recorded behavior while the other did not). The total number of agreements for the session are then divided by agreements plus disagreements and multiplied by 100 to yield the mean reliability for the entire observation.

Unfortunately, interval-by-interval calculations are sometimes impractical or impossible. This is the case if the data collection system does not permit breaking the records into smaller units. For example, assume a teacher collects data on the number of times a student raises his hand throughout a class by making tally marks on a piece of paper. For some classes, a second observer (for instance, a behavioral consultant) also records instances of hand raising using tallies. In this case, interval-by-interval reliability would be difficult to calculate because the records cannot be easily broken into smaller units; it is impossible to tell when the teacher recorded the first instance of hand raising and compare that to the consultant's data.

Also, interval-by-interval reliability methods sometimes inflate or deflate agreement based on whether behavior occurs at a high or low rate. Going back to the extreme example of an observer falling asleep, high agreement scores might occur due to the fact that not much behavior occurred. Similarly, with high rate behavior, one observer could essentially stop watching but continue to score lots of behavior and obtain a high score. To address these possibilities, it is possible to score agreement on the occurrence only intervals (i.e., evaluating only those intervals in which one observer or the other scored the occurrence of behavior) and to score agreement on the nonoccurrence intervals (i.e., evaluating only those intervals in which one observer or the other scored the nonoccurrence of behavior). By evaluating occurrence and nonoccurrence agreement, the reliability scores become less sensitive to rate fluctuations.

Some Suggestions for Integrity Measures

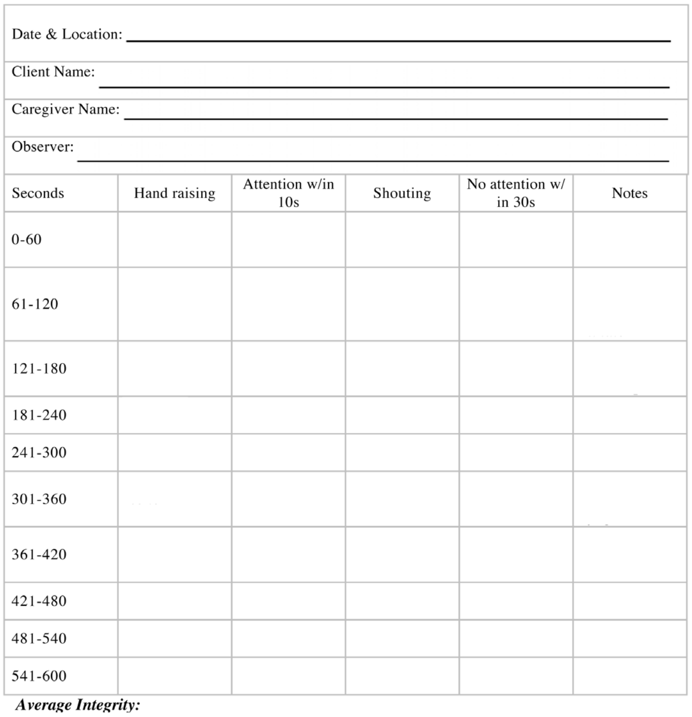
Integrity is typically calculated by examining the total percentage of opportunities for which the procedure was implemented correctly. For example, the overall integrity would be 80% if a parent correctly applied extinction to three of five undesirable responses and correctly reinforced five of five appropriate responses (eight correct responses of a total of 10 opportunities). This method is simple to explain and calculate. Unfortunately, it also lumps together different types of integrity (in this example, reinforcing appropriate behavior and failing to reinforce problem behavior); these different types of integrity may differentially affect the intervention outcome. For example, reinforcing problem behavior may be more detrimental than failing to reinforce appropriate behavior (St. Peter Pipkin, 2006). Examining integrity on individual components of an intervention may be as important as overall integrity because interventions may withstand "low" levels of integrity if the contingencies favor appropriate behavior over problem behavior. Calculating integrity measures for individual components may also allow practitioners to provide more focused feedback to caregivers. For example, if a parent consistently reinforces appropriate behavior but also periodically reinforces problem behavior, individually calculated integrity measures permit practitioners to provide specific positive and corrective feedback (respectively). This specific information is only available at a quantitative level if integrity measures are separated for each component of the intervention.

One means of calculating integrity on individual components of an intervention is to use integrity-monitoring sheets in which each component is scored individually. For example, practitioners using DRA procedures could record each instance the caregiver reinforced appropriate behavior separately from instances in which the caregiver appropriately applied extinction to problem behavior. We discuss possible data sheets for calculating integrity in this way in the next section, and provide examples in the appendices.

Suggestions for monitoring and data sheets

Monitoring reliability and integrity can be difficult for practitioners, particularly when consulting on multiple cases. To increase ease and efficiency, monitoring sessions could be relatively brief and occur on an intermittent basis. For example, practitioners could conduct a 10-min monitoring session once per week with each of their clients (e.g., Noell & Witt, 1998). There sometimes seems to be a false belief that reliability and integrity monitoring should be continuous or nearly continuous; if that were the case, the monitor himself or herself could simply conduct the procedures! Sampling is far more efficient.

During monitoring sessions, practitioners collect reliability and integrity data using data collection sheets tailored to the particular client's intervention. For example, assume a practitioner developed an intervention that involved delivering attention within 10 s of hand raising and not attending (i.e., ignoring) within 30 s of shouting. The data collection sheet for this intervention would have four sections: one each for occurrences of hand raising and shouting, one for attending following hand raising, and one for not attending following shouting. In the occurrences section, the practitioner would record the number of opportunities to implement the intervention (in this case, a count of hand raising and shouting). The number of correct caregiver responses would be recorded in the other two sections. Sample data sheets (blank and completed) are shown in Figures 2, 3, and 4. Figure 2 shows a blank data collection sheet, whereas the sheets in Figures 3 and 4 show hypothetical uses of the data sheet. In Figure 3, treatment integrity is calculated by dividing the number of correct teacher responses (delivering and withholding attention



Date & Location: ____

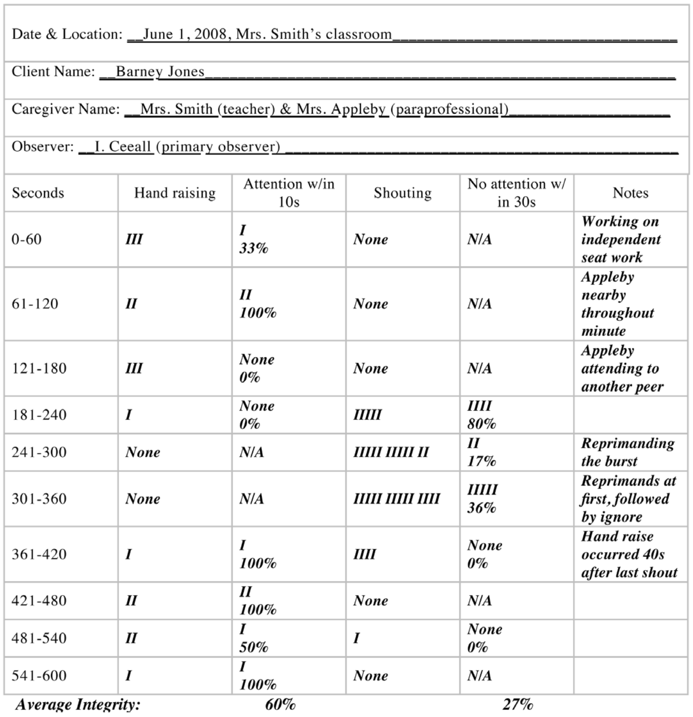
Client Name: ____

Caregiver Name: ____

Observer: ____

Seconds	Hand raising	Attention w/in 10s	Shouting	No attention w/ in 30s	Notes
0-60					
61-120					
121-180					
181-240					
241-300					
301-360					
361-420					
421-480					
481-540					
541-600					
Average Integrity:					

Figure 2. Sample blank data sheet for monitoring treatment integrity; the intervention involves delivering attention within 10 s of hand raising and not attending (i.e., ignoring) within 30 s of shouting.



Date & Location: __June 1, 2008, Mrs. Smith's classroom________________________________

Client Name: __Barney Jones____

Caregiver Name: __Mrs. Smith (teacher) & Mrs. Appleby (paraprofessional)__________________

Observer: __I. Ceeall (primary observer) ____

Seconds	Hand raising	Attention w/in 10s	Shouting	No attention w/ in 30s	Notes
0-60	***III***	***I*** ***33%***	***None***	***N/A***	***Working on independent seat work***
61-120	***II***	***II*** ***100%***	***None***	***N/A***	***Appleby nearby throughout minute***
121-180	***III***	***None*** ***0%***	***None***	***N/A***	***Appleby attending to another peer***
181-240	***I***	***None*** ***0%***	***IIIII***	***IIII*** ***80%***	
241-300	***None***	***N/A***	***IIIII IIIII II***	***II*** ***17%***	***Reprimanding the burst***
301-360	***None***	***N/A***	***IIIII IIIII IIII***	***IIIII*** ***36%***	***Reprimands at first, followed by ignore***
361-420	***I***	***I*** ***100%***	***IIII***	***None*** ***0%***	***Hand raise occurred 40s after last shout***
421-480	***II***	***II*** ***100%***	***None***	***N/A***	
481-540	***II***	***I*** ***50%***	***I***	***None*** ***0%***	
541-600	***I***	***I*** ***100%***	***None***	***N/A***	
Average Integrity:		***60%***		***27%***	

Figure 3. Sample data sheet for monitoring treatment integrity that shows hypothetical data; treatment integrity is calculated by dividing the number of correct teacher responses (delivering and withholding attention following hand raising and screaming, respectively) by the number of student responses, and multiplying by 100; overall integrity is obtained by averaging the integrity across the 1-min intervals.

following hand raising and screaming, respectively) by the number of student responses, and multiplying by 100. Integrity varies throughout the session, as shown by the integrity numbers in each block. Overall integrity is obtained by averaging the integrity across the 1-min intervals, and it is relatively low overall (60% for omission integrity and 27% for commission integrity).

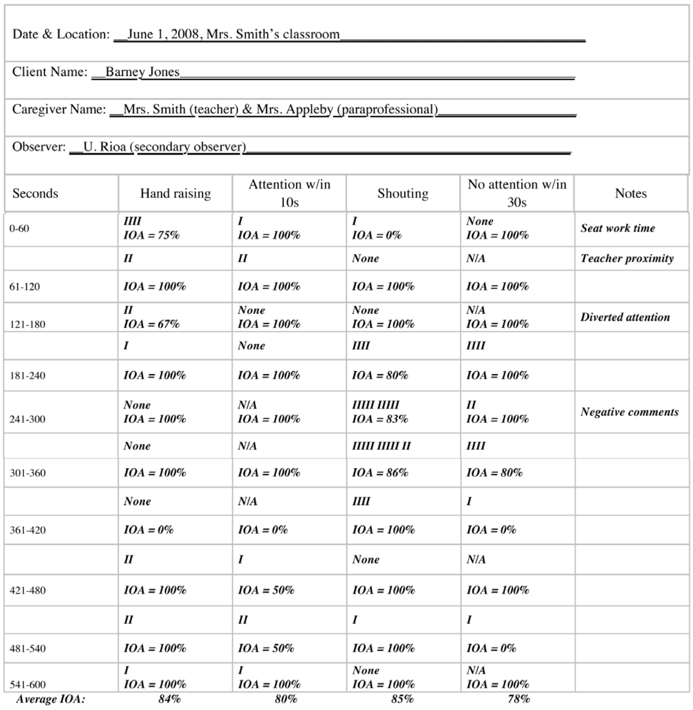
Data from brief monitoring sessions could also be used to check the reliability of caregiver data collection. Comparison of the practitioner's record with the caregiver's record could permit immediate feedback to the caregiver regarding the reliability of data collection and intervention integrity. In the example described above, the recorded "opportunities" are also the counts of hand raising and shouting. The practitioner could use a whole-session reliability measure (as described in the section on reliability measurement) by dividing the smaller number of recorded responses by the larger and multiplying by 100, or use a proportional agreement method. In Figure 4, reliability is calculated using a proportional agreement method. Agreement using this method averages between 78% and 85%. If a less stringent method of reliability calculation was more appropriate, a whole-session measure could be used, which would yield average agreement scores between 88% and 93%.

Using data sheets like these may be useful because caregivers could be immediately alerted if reliability is low. Thus, brief monitoring sessions could be conducted using relatively simple materials. Despite the simplistic data collection, these measures provide opportunities for calculating both reliability and integrity, and for providing immediate feedback to caregivers about ongoing recording of behavior and implementation of behavior-change procedures.

Conclusions

Data reliability and treatment integrity should be measured in the everyday practice of behavior analysis. Failing to do so could be dangerous, and it is nearly impossible to judge the efficacy of behavioral procedures without such data. In addition, the ability to provide feedback to data collectors and procedure implementers is paramount. Data reliability errors and treatment integrity errors can be avoided through good training, solid descriptions of definitions and procedures, generalization and maintenance training, and by making the procedures as simple and parsimonious as possible. Monitoring should also be simple and parsimonious, using efficient methods such as intermittent sampling rather than continuous monitoring.

An issue that is likely to arise for practicing behavior analysts relates to the problem of reimbursement. In short, is the practice of monitoring for reliability



Date & Location: June 1, 2008, Mrs. Smith's classroom

Client Name: Barney Jones

Caregiver Name: Mrs. Smith (teacher) & Mrs. Appleby (paraprofessional)

Observer: U. Rioa (secondary observer)

Seconds	Hand raising	Attention w/in 10s	Shouting	No attention w/in 30s	Notes
0-60	*IIII* *IOA = 75%*	*I* *IOA = 100%*	*I* *IOA = 0%*	*None* *IOA = 100%*	*Seat work time*
	II	*II*	*None*	*N/A*	*Teacher proximity*
61-120	*IOA = 100%*	*IOA = 100%*	*IOA = 100%*	*IOA = 100%*	
121-180	*II* *IOA = 67%*	*None* *IOA = 100%*	*None* *IOA = 100%*	*N/A* *IOA = 100%*	*Diverted attention*
	I	*None*	*IIII*	*IIII*	
181-240	*IOA = 100%*	*IOA = 100%*	*IOA = 80%*	*IOA = 100%*	
241-300	*None* *IOA = 100%*	*N/A* *IOA = 100%*	*IIIII IIIII* *IOA = 83%*	*II* *IOA = 100%*	*Negative comments*
	None	*N/A*	*IIIII IIIII II*	*IIII*	
301-360	*IOA = 100%*	*IOA = 100%*	*IOA = 86%*	*IOA = 80%*	
	None	*N/A*	*IIII*	*I*	
361-420	*IOA = 0%*	*IOA = 0%*	*IOA = 100%*	*IOA = 0%*	
	II	*I*	*None*	*N/A*	
421-480	*IOA = 100%*	*IOA = 50%*	*IOA = 100%*	*IOA = 100%*	
	II	*II*	*I*	*I*	
481-540	*IOA = 100%*	*IOA = 50%*	*IOA = 100%*	*IOA = 0%*	
541-600	*I* *IOA = 100%*	*I* *IOA = 100%*	*None* *IOA = 100%*	*N/A* *IOA = 100%*	
Average IOA:	*84%*	*80%*	*85%*	*78%*	

Figure 4. Sample data sheet for monitoring reliability that shows hypothetical data collected by a secondary observer; reliability is calculated using a proportional agreement method.

and integrity reimbursable? The solution to the reimbursement issue is likely to vary from state to state or country to country, or from one insurance company to another. In our view, however, the practice is as important as any other assessment or therapy component of applied behavior analysis. Hence, behavior analysts are justified in billing for their services even when, if not especially when, they are taking measures to ensure good reliability and integrity.

References

Allen, K. D., & Warzak, W. J. (2000). The problem of parental nonadherence in clinical behavior analysis: Effective treatment is not enough. *Journal of Applied Behavior Analysis, 33*, 373-391.

Brackett, L., Reid, D. H., & Green, C. W. (2007). Effects of reactivity to observations on staff performance. *Journal of Applied Behavior Analysis, 40*, 191-195.

DiGennaro, F. D., Martens, B. K., & Kleinmann, A. E. (2007). A comparison of performance feedback procedures on teachers' treatment implementation integrity and students' inappropriate behavior in special education classrooms. *Journal of Applied Behavior Analysis, 40*, 447-461.

Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S. & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review, 22*, 254-272.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10*, 103-116.

Hawkins, R. P. and Dotson, V. A. (1975). Reliability scores that delude: An Alice in Wonderland Trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp and G. Semb (Eds.), *Behavior Analysis: Areas of research and application*. Englewood Cliffs, New Jersey: Prentice-Hall.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis, 10*, 141-150.

McIntyre, L. L, Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in Journal of Applied Behavior Analysis Studies from 1991 to 2005. *Journal of Applied Behavior Analysis, 40,* 659-672.

Noell, G. H., & Witt, J. C. (1998). Toward a behavior analytic approach to consultation. In T. S. Watson and F. M. Gresham (Eds.), *Handbook of child behavior therapy*. New York, NY: Plenum Press.

Noell, G. H., Witt, J. C., LaFleur, L. H., Mortenson, B. P., Ranier, D. D., & LeVelle, J. (2000). Increasing intervention implementation in general education following consultation: A comparison of two follow-up strategies. *Journal of Applied Behavior Analysis, 33*, 271-284.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis, 8,* 43-51.

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*, 477-492.

St. Peter Pipkin, C. C. (2006). A laboratory investigation of the effects of treatment integrity failures on differential reinforcement procedures. Unpublished doctoral dissertation, University of Florida, Gainesville.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10*, 349-367.

Sulzer-Azaroff, B., & Mayer, G. R. (1991). *Behavior analysis for lasting change*. Fort Worth: Holt, Rinehart, & Winston.

Vollmer, T. R., Marcus, B. A., & LeBlanc, L. (1994). Treatment of self-injury and hand mouthing following inconclusive functional analysis. *Journal of Applied Behavior Analysis, 27*, 331-344.

Author Note

Address correspondence to Timothy R. Vollmer, Psychology Department, University of Florida, 32611 (email: vollmera@ufl.edu).