Building Consensus to Establish Expert Ratings for Direct Behavior Rating

Austin H. Johnson[1], Rose Jaffery[1], Sayward S. Harrison[2], Ajlana Music[2], Sandra M. Chafouleas[1], T. Chris Riley-Tillman[2], & Ted J. Christ[3]

University of Connecticut[1], East Carolina University[2], University of Minnesota[3]

Introduction

Direct Behavior Rating (DBR) is an efficient and technically sound behavioral assessment method that involves making a brief rating of a target behavior immediately following a pre-specified observation period (Chafouleas, Riley-Tillman, & Christ, 2009). Historically, rater accuracy using DBR has been evaluated by comparing DBR-derived data to "true scores" yielded from systematic direct observation (SDO). However, given the fundamental differences between the two methodologies, this comparison may not be appropriate. The purpose of this study was to establish and evaluate true score estimates using expert-completed DBR.

Method

Participants. A total of 13 professors and doctoral students across two sites participated as subject-matter experts (SMEs) during the expert-rating procedure. Three doctoral students in school psychology (two at one site, one at the other) acted as session facilitators to set up and oversee the expert-rating session.

Materials. During each expert-rating session, video clips were displayed using a computer and projector. An easel, large paper, and markers were used to record and display participants' DBR scores during discussions of each video clip. Each SME was given a pen, a rating packet with a separate page for each video clip to be rated, a sheet with operational definitions of the target behaviors, and the DBR Wording Preference Questionnaire.

Video clips. Videos for rating consisted of eighteen 1-minute video clips depicting a simulated elementary-level classroom. Prior to each clip, participants were instructed to focus their attention on a particular child in order to rate one of three target behaviors: *respectful* (RS), *academically engaged* (AE), or *disruptive* (DB). After each clip, a blank screen was displayed for 30 seconds in order to allow time for participants to make their ratings. The video clips were purposely selected to include desired behavior levels (low, medium, or high) and to ensure that clips were balanced by gender of the target student. The sequence in which video clips were displayed was randomly ordered.

Procedure. Each consensus-building session lasted approximately two hours. First, SMEs viewed the first section of an online DBR Training Module as a group (providing an overview of DBR and the three behaviors to be evaluated), as well as three initial video clips of simulated elementary-level classroom instruction to practice the consensus-building procedure and clarify any initial questions. After all discussion of these practice clips ended, the SMEs viewed and rated the first nine of the 18 video clips officially targeted for evaluation. After all nine clips were rated, the session facilitators collected each rating packet and recorded the ratings for each clip onto the easel for public display. Ratings were anonymously recorded to minimize peer influence. The SMEs were instructed to discuss the ratings for each clip, particularly if there were substantial disagreements. The same nine clips were then viewed and rated a second time in order to provide final ratings for these clips in the Rating Packet. Next, the SMEs viewed the last nine target video clips and repeated this procedure.

Procedural Flowchart

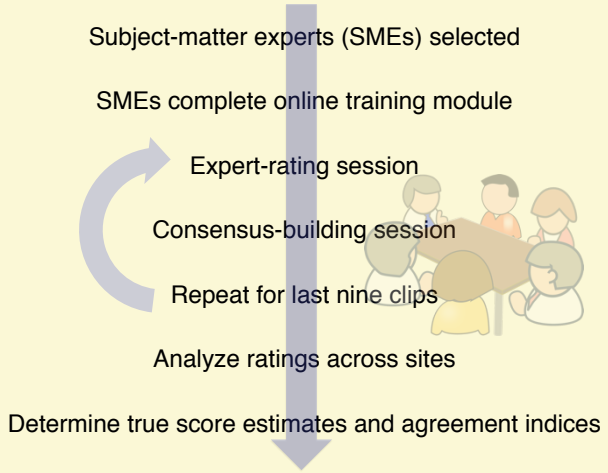

Table 1. Subject Matter Expert Agreement via r_{WG} and ICC(K).

			SME Source					
			Site 1		Site 2		Aggregated	
Clip	Behavior	Level	Initial r_{WG}	Final r_{WG}	Initial r_{WG}	Final r_{WG}	Initial r_{WG}	Final r_{WG}
1	RS	Med.	.81	.94	.80	.83	.69	.83
2	AE	Low	.92	.99	.97	.97	.94	.98
3	DB	Low	.97	.97	.97	.97	.97	.97
4	RS	High	.99	.99	.98	.98	.99	.99
5	RS	Low	.80	.95	.65	.77	.71	.82
6	DB	Med.	.94	.97	.95	.95	.94	.96
7	AE	High	1.00	1.00	1.00	1.00	1.00	1.00
8	AE	Med.	.16	.84	.35	.83	.27	.85
9	DB	High	.99	.99	.97	.97	.98	.98
10	RS	High	.54	.97	.78	.95	.65	.96
11	AE	Low	.42	.94	.97	.97	.64	.96
12	AE	Med.	.89	.97	.57	.87	.79	.93
13	DB	Low	1.00	1.00	1.00	1.00	1.00	1.00
14	RS	Low	.97	.97	.92	.92	.94	.94
15	DB	Med.	.44	.84	.78	.93	.59	.88
16	DB	High	.89	.97	.90	.90	.90	.94
17	AE	High	.98	.98	.97	.97	.97	.97
18	RS	Med.	.97	.97	.87	.98	.94	.94
M			.82	.96	.86	.93	.83	.94
SD			.25	.05	.18	.07	.20	.06
ICC(K)			.984	.997	.978	.989	.990	.997

Note. AE = Academically Engaged. DB = Disruptive. RS = Respectful.

Results

True score estimates. The median scores were chosen to serve as the "true" expert DBR-SIS scores for three reasons. First, the differences between the mean and median scores for each clip were small (range: 0 – 0.54) in comparison to the 11-point range of the DBR scale utilized in this study; therefore, choosing the median score over the mean did not considerably affect the final true score estimate. Second, the median is the measure of central tendency least affected by outliers. Finally, as these scores will be utilized as criteria for other DBR-SIS scores, they should represent values that can actually be achieved on an 11-point DBR scale (i.e., whole numbers). Mean and median values are presented in Table 2.

Agreement indices. In order to determine the level of agreement among ratings both within and across sites, two agreement indices from the industrial/organizational psychology literature were employed: r_{WG} (James, Demaree, & Wolf, 1984; 1993) and McGraw and Wong's (1996) ICC(K). Both are interpreted on a scale from 0 to 1.0, and are presented in Table 1. The r_{WG} index represents the proportion of non-error variance in ratings. ICC(K) provides an indication of the absolute consensus among raters regarding both rater consensus and relative consistency.

To determine if agreement increased after consensus-building procedures were implemented, a paired-samples t-test was conducted using the mean r_{WG} values for the initial and final ratings for each clip across both sites. Results of this analysis suggested that agreement significantly increased after experts engaged in consensus-building procedures, $t(17) = 2.87$, $p < .05$, $d = .76$.

Table 2. True Score Estimates.

Clip	*M*	*Mdn*	|*M* - *Mdn*|
1	2.69	3	0.31
2	1.31	1	0.31
3	1.62	2	0.38
4	9.85	10	0.15
5	1.38	1	0.38
6	7.38	7	0.38
7	10.00	10	0.00
8	3.85	4	0.15
9	9.23	9	0.23
10	7.38	7	0.38
11	0.54	0	0.54
12	8.08	8	0.08
13	0.00	0	0.00
14	0.85	1	0.15
15	6.77	7	0.23
16	9.31	9	0.31
17	9.62	10	0.38
18	1.62	2	0.38

Summary and Conclusions

The current study employed expert-rater methodologies from the industrial/organizational psychology literature to determine true score estimates for behavior assessment purposes. Even before consensus-building procedures were employed, initial ratings possessed high levels of agreement for most clips. For those that lacked initially high levels of agreement, no consistent pattern emerged, as all levels and types of behavior used in this study were represented among the four clips for each site that possessed low initial agreement. Furthermore, only two of the four clips (clips 8 and 10) demonstrated poor agreement for both sites.

After consensus-building procedures were implemented, a statistically significant increase was observed in the aggregate agreement across all clips and sites, indicating that this procedure is effective for increasing agreement levels.

Preparation of this poster was supported by a grant from the Institute for Education Sciences (IES), U.S. Department of Education (R324B060014). For additional information, please direct all correspondence to Sandra Chafouleas at sandra.chafouleas@uconn.edu