Cognitive Therapy and Research, Vol. 26, No. 1, February 2002 (© 2002), pp. 135–142

Interobserver Reliability of Descriptive Experience Sampling

Russell T. Hurlburt[1,2] **and Christopher L. Heavey**[1]

The descriptive experience sampling (DES) method is a procedure for providing descriptions of the inner experience of individuals. Although analytic arguments for its reliability are found, there have been no conventional interobserver reliability studies of the method. We took a stratified random sample of 10 participants and obtained 6 DES samples from each. Two interviewers independently interviewed the participants and rated them on the presence or absence of 16 characteristics of inner experience, 5 of which occurred frequently enough to analyze separately. The single-sample interobserver-reliability kappas for those 5 characteristics ranged from .52 to .92 (median 0.76). Spearman–Brown adjustment showed that reliabilities for typical 19-sample averages would range from .92 to .98, comparable to highly reliable questionnaires.

KEY WORDS: descriptive experience sampling; inner experience; reliability; cognition.

The purpose of this study was to explore the reliability of the descriptive experience sampling (DES) method when used as a procedure that rates predefined characteristics. DES (Hurlburt, 1990, 1993) is a method designed to describe the characteristics of awareness. The method has four steps: (1) the participant carries a random beeper in his or her natural environment. When the randomly occurring beep sounds, the participant is to pay particular attention to the awareness that was ongoing at the moment of the beep and to jot down in a notebook enough notes about that momentary experience to be able to describe it during a subsequent interview. A series of five-to-eight such samples are typically collected on any given sampling day. (2) Later that same day or the next day, the investigator interviews the participant about those five-to-eight sampled moments to obtain as complete an understanding of those moments as possible, and then writes a description of each momentary awareness. (3) After obtaining a series of such descriptions across several (4–8) sampling days, the investigator reviews the entire set of descriptions and extracts the salient characteristics—features of awareness that occurred in several

[1]Department of Psychology, University of Nevada, Las Vegas, Nevada.
[2]Correspondence should be directed to Russell T. Hurlburt, Department of Psychology, University of Nevada, Las Vegas, Nevada 89154-5030; e-mail: russ@unlv.edu.

0147-5916/02/0200-0135/0

or many of the sampled moments. Depending on the purpose of the study, a fourth step is often included: (4) After several participants, all of whom share some external characteristic (e.g. psychiatric diagnosis), have undergone Steps 1 through 3, the investigator extracts the common characteristics—features of awareness that occur in several or many of the participants' salient characteristics that had been extracted in Step 3 for each participant.

Thus the ultimate goal of DES is first to describe the characteristics of the awareness of one individual and then, often, to identify the features of awareness that some group of individuals share in common. Hurlburt (1990, 1993, 1997) showed that the DES method distinguishes between psychiatric categories, and Hurlburt, Koch, and Heavey (this issue) showed that the method distinguishes between nondiagnosable individuals who share an easily observable external characteristic (high speech rate).

Hurlburt (1993, 1997) contended that the method's ability to discriminate between psychiatric categories is evidence for its validity and hence for its reliability. He also provided case studies and argued for their reliability. However, those arguments may apply only to the specific cases involved.

Thus there have been several analytic arguments in support of the reliability of DES but no conventional reliability study. The natural environment, sampling, and interview features of DES present challenges not found in conventional reliability studies. Because the method is new, there is no immediate context for its reliability measurement. The reliability studies performed on the structured clinical interview for *DSM-III* and *DSM-IV* (SCID) provide an analogous context because both DES and SCID involve interviews whose aim is rating the presence or absence of characteristics (psychiatric diagnoses in the case of the SCID). Interobserver reliability of the SCID has been assessed in the following ways: test-retest after a short interval (Dreeson & Arntz, 1998; First et al., 1995; Williams et al., 1992); test-retest where the initial interview is live and the retest raters watch a videotape (Riskind, Beck, Berchick, Brown, & Steer, 1987) or listen to an audiotape (Renneberg, Chambless, Dowdall, Fauerbach, & Gracely, 1992) of the original interview; test-retest where both raters watch the same videotaped interview (Ventura, Liberman, Green, Shaner, & Mintz, 1998); and joint interviews (Arntz et al., 1992; Brooks, Baltazar, McDowell, Munjack, & Bruns, 1991; Renneberg et al., 1992; Stanley, Turner, & Borden, 1990; Wonderlich, Swift, Slotnick, & Goodman, 1990).

Of these, the short-interval test-retest method is the most conservative way of assessing SCID reliability because there are more sources of variance.

> The short-interval test-retest approach is a more thorough examination of reliability. In this approach, two raters separately interview the same patient. In this way, three sources of variance are being tested: (1) rater variance in the elicitation of information, (2) rater variance in the interpretation of information, and (3) patient variance in providing information across interviews. In the joint-interview approach, only rater variance in the interpretation of information is being tested. . . . Compared to joint-interview reliabilities, short interval test-retest reliabilities are . . . more generalizable to the actual practice, in which different raters conduct the interview. (Dreesen & Arntz, 1998, p. 139)

Taped interviews have the same source of variance as joint interviews, namely rater variance in interpretation of information; Dreesen and Arntz conclude that short-interval test-retest is the most thorough and most generalizable way of assessing the

reliability of the SCID. Given that this conclusion applies equally to DES, we chose this design for this study.

To assess the interobserver reliability of DES, we had two investigators independently interview and subsequently rate the same series of sampled moments from a group of participants. This design answers two important questions: (1) To what extent do two independent interviewers arrive at the same DES ratings? (2) Can DES be applied by anyone other than its originator (the senior author)?

Hurlburt (1990, 1993) identified 16 characteristics of experience, each of which had been idiographically observed to occur across several or many participants: inner speech, partially worded speech, unworded speech, worded thinking, image, imageless seeing, unsymbolized thinking, inner hearing, feeling, sensory awareness, just doing, just talking, just listening, just reading, just watching TV, and multiple awareness. For this study we created a codebook that contained descriptions of these 16 characteristics along with discussions of relevant rating considerations. We pretested this codebook with 2 participants and made modifications accordingly. The complete codebook is available on the World Wide Web (Hurlburt & Heavey, 2000).

The study had two phases. Our aim in Phase I was to obtain a heterogeneous sample of inner experiences that was at least approximately representative of the range of naturally occurring inner experiences. Our aim in Phase II was to have two independent interviewers, including one skeptical nonoriginator (the junior author), use the DES procedure and codebook to rate each participant's samples. We then computed reliability coefficients for these two sets of ratings.

METHOD

Phase I

Participants

All 210 students in attendance at one lecture meeting of an introductory psychology class were participants in Phase I and received course research-participation credit.

Materials and Apparatus

The Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1994) was administered to all participants. The SCL-90-R asks respondents to indicate the extent to which they are bothered by each of 90 symptoms on a scale from 0 (*not at all*) to 4 (*extremely*). The Global Severity Index (GSI), used in this study, is the sum of all responses and indicates the overall extent of symptom distress. All participants also filled out a brief demographic questionnaire that asked their age and gender as well as other questions relevant to another study but not used here.

Procedure

The investigators gave a brief description of this study to the 210 students attending an introductory psychology class and invited them to fill out the SCL-90-R, the

demographic questionnaire, and a consent form in exchange for research-participation credit and the opportunity to proceed to Phase II; all students did so.

Phase II

Participants

The 210 participants from Phase I were stratified into 10 equal-sized strata (21 per stratum) on the basis of their SCL-90-R GSI scores. One prospective participant from each stratum was selected at random and invited to participate in Phase II. If a prospective participant declined to participate, he or she was replaced by another randomly chosen participant from the same stratum; this replacement occurred a total of four times because of scheduling conflicts (3) and lack of interest (1). Sex of participant was alternated between each stratum. These Phase II participants' ages ranged from 18 to 19, and they received additional course research-participation credit and $30 as described below. The two interviewers participating in this study were this paper's two authors (the second author was a skeptic who was not the originator of the method).

Materials and Apparatus

A portable shirt-pocket-sized beeper was used by each participant in Phase II. The beepers (Hurlburt, 2000c) emitted a 700-Hz beep through a transistor-radio-type earphone at random intervals whose mean length was 30 min (maximum 60 min). Participants in Phase II were also supplied with a 3 in. × 5 in. notebook.

Procedure

Each participant in Phase II participated in four meetings: the explanation meeting and sampling meetings 1, 2, and 3. They received $5 at the end of sampling meetings 1 and 2 and $20 at the end of sampling meeting 3. There were no dropouts and no missing data. The interviewers and participants were kept blind to the participants' SCL-90-R scores and strata information throughout the study.

In the explanation meeting, one of the interviewers (randomly assigned) explained the study and its procedure, obtained informed consent to continue, instructed the participant in the DES method, and scheduled the three sampling meetings (typically spread over 1–2 weeks). Participants were instructed to wear the beeper in their usual, everyday environments until they had responded to five beeps (which typically required 2–3 hr) on the day of or the day before the next sampling meeting. They were instructed to "take a mental snapshot" of whatever was occurring in their awareness at the very moment of the onset of each beep—that last microsecond undisturbed by the beep itself—and to jot down enough notes to be able to discuss that particular moment in detail in the sampling interviews. They were told that we were interested simply in a complete, accurate description of their awareness as it happened to exist at the moment of the beep. We defined awareness as broadly as possible, including thoughts, feelings, sensations, perceptions, tickles, and so on. We said we were not interested in whether those awarenesses were typical

or atypical, or in potential explanations of the source of those awarenesses. All participants were advised that they should feel free to decline to describe any sampled experience, and that we would maintain their confidentiality. We treated participants as coresearchers and encouraged them to take that role, explaining that while we had a method that enabled us to examine inner experience in detail, they had the inner experience itself. Together we might discover something that neither of us could alone.

Each sampling meeting was divided into two halves, one with each of the interviewers separately. The order of interviewers was counterbalanced and alternating. Each interviewer discussed the samples that the participant had obtained and scored the first three of them according to the presence or absence of each of the 16 characteristics described in the codebook. We had asked the participants to supply five samples so that if they did not comply completely, or if some of the samples were unusable, we would have three samples from each day; in fact, we were able to use the first three samples for each participant. The participant was instructed to answer each interviewer's questions but not to discuss the contents of the other interviewer's interview. The participant was not informed that his or her descriptions would be rated. The first sampling meeting was considered part of the participants' training process; ratings from this meeting were not included in the analyses. Therefore, each of the 10 participants contributed six samples to the analyses.

RESULTS AND DISCUSSION

The 10 participants included five men and five women with a mean age of 18.3 (range 18–19). The mean SCL-90-R GSI score was 65.10 ($SD = 45.03$). As expected because of the stratification, this closely mirrored the GSI distribution of the entire Phase I group ($M = 69.82$, $SD = 46.45$).

Each interviewer rated each of the $10 \times 6 = 60$ samples as to the presence or absence of the 16 characteristics listed in the codebook. Of those 16 characteristics, five (images, inner speech, unsymbolized thinking, feelings, and sensory awareness) were rated by both interviewers as occurring in at least 15 of the 60 samples. The remaining 11 codebook characteristics were rated as occurring in five or fewer total samples. We therefore focus our analysis on the five frequently occurring characteristics and exclude the remaining 11 low frequency characteristics because the reliability statistics may be unstable. This follows Dreesen and Arntz's analysis of SCID reliability (Dreesen & Arntz, 1998) and is a conservative step because including them would make the DES reliability stronger: The percentage of interobserver agreement on the aggregate of the 11 low frequency characteristics was somewhat higher (98.8%) than that on the aggregate of the five frequently occurring characteristics (91.3%).

For each characteristic, there are three ways that interobserver reliability can be assessed: A "samplewise" analysis considers the interobserver agreement on particular samples; a "participantwise" analysis considers the interobserver agreement on the average frequency of occurrence of the characteristic across all six of the participant's samples; and a "typical" analysis extrapolates the participantwise six-sample analysis to a more typical DES study. We will discuss each in turn.

Table I. Frequencies and Reliability Statistics for Five Characteristics

Characteristic	Frequency (%)	Samplewise: Observer agreement (%)	Samplewise: Interobserver reliability (Kappa)	Participantwise reliability of 6-sample average (Pearson *r*)	Typical reliability[a] of 19-sample average (Pearson *r*)
Images	32	97	.92	.95	.98
Inner speech	31	95	.88	.84	.94
Unsymbolized thinking	21	90	.69	.95	.98
Feelings	24	92	.76	.93	.98
Sensory awareness	22	83	.52	.78	.91

[a] Spearman–Brown estimates.

For each of the five characteristics, samplewise reliability considers the extent to which observers agreed when rating the 60 particular samples. The percentage of agreements ranged from 83 to 97% and are shown in the second column of Table I. Kappa statistics ranged from .52 to .92 and are shown in the third column of Table I. Thus at the level of single samples, there is a substantial agreement between the two interviewers: The two interviewers agreed on 91% of all ratings made for these five characteristics, and the median kappa was 0.76. These interobserver (samplewise) reliability coefficients compare quite favorably to the analogous single-item reliabilities generally found in questionnaires and are substantially higher than those generally found in SCID studies.

Participantwise reliability is analogous to a questionnaire's scale reliability. For each characteristic, we computed the participantwise score for each interviewer by averaging all six of that interviewer's ratings of the participant. We then computed the participantwise reliability by correlating those six-sample averages between interviewers across the 10 participants. Those correlations are shown in the second-to-last column of Table I. The median of the participantwise (six-sample average) reliabilities was 0.93.

Typical DES studies use more than six samples per participant. For example, Hurlburt, Koch, and Heavey (in press) obtained between 16 and 22 samples ($M = 19$) from each participant in their target group. If for the sake of argument we accept 19 samples as a representative number of samples in a typical DES study, we can apply the Spearman–Brown formula (Anastasi, 1988) to estimate the typical 19-sample-average reliability of DES characteristics, as shown in the last column of Table I. The median of these Spearman–Brown estimates of the reliability of 19-sample averages is 0.98.

The characteristic with the lowest reliability was sensory awareness, primarily reflecting the difficulty in determining whether a bodily awareness was or was not part of a feeling. For example, if a participant experienced anxiety that was manifested as a pain in his chest, this was scored as a feeling, not a sensory awareness. However, if the participant was particularly noticing the characteristics of a pain in his chest, then this was scored a sensory awareness regardless of whether the pain was caused by a feeling. This kind of discrimination proved somewhat difficult, resulting in a kappa of 0.52. Even so, the typical 19-sample reliability for sensory awareness was .91.

Table II. Item Interdependencies (Kappa Statistics) for Each Rater

	Images	Inner speech	Unsymbolized thinking	Feelings	Sensory awareness
Images	.92	−.27	−.30	−.20	−.17
Inner speech	−.37	.88	−.28	−.08	−.05
Unsymbolized T.	−.26	−.25	.69	−.06	.04
Feelings	−.09	.09	−.08	.76	−.22
Sensory A.	−.19	−.17	−.15	.03	.52

Note. Interdependency kappas for interviewer 1 are given below the diagonal and for interviewer 2 above the diagonal, and interobserver kappas (same as Table I column 3) are on the main diagonal for comparison.

Thus we have seen that the five most frequently occurring DES characteristics can be rated reliably. We investigated the extent to which these characteristics are distinct entities by computing kappas for all possible pairs of the five characteristics for each interviewer across all 60 samples. Table II shows these interdependency kappas for interviewer 1 below the main diagonal and for interviewer 2 above the main diagonal. For ease of comparison, the main diagonal of Table II redisplays the interobserver reliabilities from the third column in Table I. Inspection of Table II reveals that the (off-diagonal) intercharacteristic kappas are in fact much smaller than the (main diagonal) interobserver reliability kappas, as they should be if the characteristics are distinct.

The establishment of high levels of reliability does not, of course, imply that the participants in this study were necessarily reporting accurately about the characteristics of their inner experience—the same reliability coefficients might have been obtained if all participants had been lying consistently. Although technically true and fundamentally irrefutable, such a claim is emphatically denied by all participants, who during debriefing said they had been accurately reporting their experiences.

This design has the potential weakness of "leakage" from one interview to the next: The participant's reports to the second interviewer may to some extent be influenced by the first interview. However, it should be noted that the participant is asked simply to *describe* his or her own inner experience, not to *rate* it—the rating task is performed independently by the observers. In fact, the participants were unaware that any rating was being performed. Because for leakage to occur in this study the participant must actively transmit the "leak" from one interviewer to the next, this methodological weakness is less problematic here than in joint- or taped-interview methods, where any leak is directly observable by both raters. Leakage is not generally considered problematic in SCID reliability studies, most of which use joint- or taped-interview method.

All methods of evaluating DES reliability contain some methodological flaws. That does not nullify DES as a method but requires that different investigators ascertain reliability in different ways to see if reliability estimates converge.

This study lends strong support to the notion that independent observers can rate reports of inner processes consistently. Furthermore, because the junior author was skeptical about the method at the outset of this study, it demonstrates that the method can be applied by someone other than its originator (the senior author).

Thus far this technique has been applied only by Hurlburt and colleagues, and so independent investigations by other laboratories are clearly called for.

ACKNOWLEDGMENTS

We thank Michael Thorsteinson and Lenard Peterson for their help with this study and Marta Meana for comments on earlier drafts of this paper.

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Arntz, A., van Beijsterveldt, B., Hoekstra, R., Hofman, A., Eussen, M., & Sallaerts, S. (1992). The interrater reliability of a Dutch version of the Structured Clinical Interview for DSM-III-R personality disorders. *Acta Psychiatrica Scandinavica*, *85,* 394–400.

Brooks, R. B., Baltazar, P. L., McDowell, D. E., Munjack, D. J., & Bruns, J. R. (1991). Personality disorders co-occurring with panic disorder with agoraphobia. *Journal of Personality Disorders*, *5,* 328–336.

Derogatis, L. R. (1994). *The SCL-90-R: Scoring, administration, and procedures* (3rd ed.). Minneapolis, MN: National Computer Systems.

Dreesen, L., & Arntz, A. (1998). Short-interval test-retest interrater reliability of the Structured Clinical Interview for DSM-III-R personality disorders (SCID-II) in outpatients. *Journal of Personality Disorders*, *12,* 138–148.

First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., Davies, M., Borus, J., et al. (1995). The Structured Clinical Interview for DSM-III-R personality disorders (SCID-II). Part II: Multi-site test-retest reliability study. *Journal of Personality Disorders*, *9,* 92–104.

Hurlburt, R. T. (1990). *Sampling normal and schizophrenic inner experience*. New York: Plenum.

Hurlburt, R. T. (1993). *Sampling inner experience in disturbed affect*. New York: Plenum.

Hurlburt, R. T. (1997). Randomly sampling thinking in the natural environment. *Journal of Consulting and Clinical Psychology*, *65,* 941–949.

Hurlburt, R. T. (2000a). *Denizens of the phenom: The features of awareness*. Manuscript submitted for publication.

Hurlburt, R. T. (2000b). *Should we believe Descriptive Experience Sampling results? Transcript of a sampling interview*. Manuscript submitted for publication.

Hurlburt, R. T. (2000c). *V.3.x random interval generator ("beeper")*. Retrieved January 16, 2000, from University of Nevada, Las Vegas web site: http://www.nevada.edu/~russ/beeper.html

Hurlburt, R. T., & Heavey, C. L. (2000). *Descriptive Experience Sampling manual of terminology*. Retrieved January 16, 2000, from University of Nevada, Las Vegas Web site: http://www.nevada.edu/~russ/codebook.html

Hurlburt, R. T., Koch, M., & Heavey, C. L. (2002). Descriptive Experience Sampling can demonstrate the connection of thinking to externally observable behavior. *Cognitive Therapy and Research*, *26,* 117–134.

Renneberg, B., Chambless, D. L., Dowdall, D. J., Fauerbach, J. A., & Gracely, E. J. (1992). The Structured Clinical Interview for DSM-III-R, Axis II and the Millon Clinical Multiaxial Inventory: A concurrent validity study of personality disorders among anxious outpatients. *Journal of Personality Disorders*, *6,* 117–124.

Riskind, J. H., Beck, A, T., Berchick, R. J., Brown, G., & Steer, R. A. (1987). Reliability of DSM-III diagnoses for major depression and generalized anxiety disorder using the Structured Clinical Interview of DSM-III. *Archives of General Psychiatry*, *44,* 817–820.

Stanley, M. A., Turner, S. A., & Borden, J. W. (1990). Schizotypal features in obsessive–compulsive disorder. *Comprehensive Psychiatry*, *31,* 511–518.

Ventura, J., Liberman, R. P., Green, M. F., Shaner, A., & Mintz, J. (1998). Training and quality assurance with the Structured Clinical Interview for DSM-IV (SCID-I/P). *Psychiatry Research*, *79,* 163–173.

Williams, J. B. W., Gibbon, M., First, M. B., Spitzer, R. L., Davies, M., Borus, J., et al. (1992). The Structured Clinical Interview for DSM-III-R (SCID). II: Multisite test-retest reliability. *Archives of General Psychiatry*, *49,* 630–636.

Wonderlich, S. A., Swift, W. J., Slotnick, H. B., & Goodman, S. (1990). DSM-III-R personality disorders in eating-disorder subtypes. *International Journal of Eating Disorders*, *9,* 607–616.