



PERGAMON

Journal of Behavior Therapy
and Experimental Psychiatry 30 (1999) 63–69

JOURNAL OF
behavior
therapy and
experimental
psychiatry

Using item analysis to facilitate interpretation of empirical findings

Kenneth J. Ruggiero*, Jeffrey L. Goodie, Tracy L. Morris

Department of Psychology, West Virginia University, P.O. Box 6040, Morgantown, WV 26506-6040, USA

Abstract

Researchers often present and interpret empirical findings with reference to hypothetical constructs and diagnostic labels. Such interpretations commonly are based upon “summary” scores obtained through interview, self-report, or rating-scale assessment instruments. Although there are advantages associated with communicating empirical findings through analysis with summary scores, there also are weaknesses that may limit the interpretability of empirical findings and impede theory development. We discuss the importance of item analysis as a tool that may guide presentation of empirical findings, and we describe how it may be used to minimize these limitations of assessment, facilitate data interpretation, and increase the opportunity for theoretical advances. © 1999 Elsevier Science Ltd. All rights reserved.

1. Introduction

Behavioral classification, long considered a useful tool for researchers and clinicians, involves the development of “summary” labels that provide a useful starting point from which to examine relations among clusters of covarying behaviors (Adams & Cassidy, 1993; Scotti, Morris, McNeil, & Hawkins, 1996). Labels derived through behavioral classification also are useful because they enable clinicians to communicate efficiently with other professionals about a client’s presenting complaint(s), and often are used by researchers to facilitate presentation and interpretation of empirical findings. One example of a widely used behavioral classification system is the *Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 1994)*. The *DSM* serves many purposes for researchers and clinicians, as it depicts an extensive range of ever-evolving diagnostic categories employed in research and clinical practice.

*Corresponding author. Tel.: 304 293-2001/ext 891; Fax: 304 293-6606; E-mail: kruggier@wvu.edu.

Despite the many advantages of behavioral classification, researchers cannot fully illustrate empirical findings with reference to diagnostic labels (e.g., “posttraumatic stress disorder” (PTSD)) or hypothetical constructs (e.g., “anxiety”, “aggressiveness”) unless they also examine *individual symptom endorsements* that pertain to these labels and constructs. Thus, when exploring diagnostic prevalence for a defined group of individuals (e.g., PTSD prevalence for individuals exposed to a specific traumatic event), it also may be important to include an examination of *symptom prevalence*. Similarly, when examining construct-related assessment scores (e.g., scores obtained on a measure of depression), it also may be important to examine responses to individual items comprising the instrument. By supplementing global research findings with item analysis results, researchers may provide clinically and theoretically relevant information on both a molar (e.g., diagnostic) and molecular (e.g., symptom prevalence) level. We will discuss how this might be accomplished, but first will explore two methods of assessment that often are employed to derive labels or scores that serve to summarize symptom endorsements: structured interviews and rating-scale instruments.

2. The role of “indirect” assessment in behavioral classification

Interviews and rating-scale assessments typically involve retrospective verbal depictions of behavior rather than direct observations, and therefore are referred to as “indirect” methods of assessment (Cone, 1978). Interviews commonly are used to facilitate diagnostic labeling, as many are designed to coincide with diagnostic criteria delineated in the *DSM* (e.g., Schedule for Affective Disorders and Schizophrenia for School-Age Children: Present and Lifetime Version; Kaufman et al., 1997; Anxiety Disorders Interview Schedule for *DSM-IV*, Brown et al., 1994). Similarly, rating-scale assessments often are used to assess reported symptomatology associated with a particular construct, such as anxiety or depression.

When using interview and rating-scale assessment procedures, participant-endorsed symptomatology may be “summarized” using scores or labels that are based on collective responses to assessment items. For instance, responses to items on the Children’s Depression Inventory (CDI; Kovacs, 1992) may be summed to provide a total score that represents the “severity” of depression (e.g., some individuals recognize cutoffs purported to represent “mild”, “moderate”, or “severe” depression). Although interviews usually are not designed to yield summary scores such as those found with rating-scale instruments, summary labels (e.g., *DSM* diagnoses) are derived when symptom endorsements satisfy the diagnostic criteria for that label. Summary labels and summary scores may be used to facilitate communication of information in similar ways. For instance, just as it is useful to communicate that 40% of a sample of individuals met criteria for “generalized anxiety disorder”, it similarly may be useful to describe that 40% of a sample of individuals obtained scores of 20 or greater on a scale of “trait anxiety”.

Although summary scores and labels derived through indirect methods of assessment have considerable utility in research and clinical practice, they also have several

limitations that imply a need for clarification beyond these scores and labels, particularly when presenting and interpreting empirical findings. One weakness of summary scores and labels involves the limited extent to which researchers may use them to describe or elucidate behavioral relations (Sprock & Blashfield, 1983), as evidenced by potentially-significant discrepancies in behavioral presentations among individuals with identical diagnostic labels. To illustrate, the *DSM* diagnosis “conduct disorder” may be used to describe a 7-year-old child who threatens other children, initiates fights, and hurts animals, but also may be used to describe a 15-year-old adolescent who sets fires, has committed armed robbery, and has forced someone into sexual activity. A second limitation associated with summary scores and labels involves the ever-changing criteria that characterize constructs and diagnostic categories psychologists employ. For instance, it may be of little utility to future readers if one concludes that a particular variable (e.g., a specific parental discipline technique) is related to the development of “conduct disorder” if the diagnostic criteria for conduct disorder are revised significantly in future *DSM* publications. Finally, a third limitation surrounding the use of summary scores and labels involves the overlapping features found within many constructs and diagnostic categories. For instance, some CDI items that correlate with “depression” also may correlate with other constructs or diagnostic categories, such as “oppositional defiant disorder” (e.g., “I never do what I am told”) and “conduct disorder” (e.g., “I get into fights all the time”).

These limitations accentuate the importance of clarification beyond summary information when presenting and interpreting empirical findings. One of the ways in which these limitations may be minimized involves supplementing summary scores and labels with additional information. Thus, we propose that by supplementing presentation and interpretation of summary results with analyses involving individual item responses, researchers may increase the interpretability of their findings and enable discussion pertaining to specific characteristics of the label or construct assessed.

3. Interpretation of empirical findings: the benefits of item analysis

An example of communicating important, but insufficient, information would be to report that children who experienced a specific traumatic event tend to exhibit a greater number of depressive symptoms (e.g., obtained higher scores on a measure of depression) than those who had not experienced the event. This statement may enable clear and concise communication of research findings, but the extent to which this statement may be interpreted is limited. In order to capitalize on the strengths of indirect methods of assessment, we feel it is important to supplement this information with an examination of the individual depressive symptoms (e.g., appetite, mood, sleep changes) assessed with the instrument employed in the study. Individual item responses provide a level of specificity that cannot be determined from summary scores obtained in group-design research.

In addition to offering increased interpretability of empirical findings, item analysis may facilitate theory development. For instance, theoretical advances concerning the

effects of a particular traumatic event may be more likely to occur if we are able to recognize the specific PTSD-related symptoms that are most commonly endorsed by individuals with similar traumatic histories. Item analysis also may be beneficial when presenting and interpreting empirical findings because it may promote the identification of subtypes of disorders (e.g., labels that characterize specific combinations of conduct-related behavioral presentations) and variables upon which a clinician may effectively intervene.

4. An illustrative example

Suppose we conducted a study to investigate the behavioral correlates of a traumatic event *X*, and wished to examine differences on the CDI between children who had directly experienced the event and two control-group samples (clinical and non-clinical) of children who had not experienced the event. Assuming the study design included additional assessments and procedures, many researchers may have elected to present and interpret findings based solely upon analyses with summary scores (e.g., between-group differences in CDI scores). One might report, for instance, that an analysis of variance yielded lower CDI scores for children in the non-clinical group when compared to children exposed to the traumatic event and the non-exposed clinically-referred sample. Although this statement is important and may convey information sufficient for theoretical advancement and interpretation of findings, it is likely that one could communicate a greater breadth of clinically and theoretically pertinent information by supplementing this statement with results of analyses that pertain to the individual CDI items. Thus, it may be useful to delineate (e.g., in Table Y) between-group differences for individual items comprising the CDI (see Table 1 for an illustration of how such a table might be constructed).

The *hypothetical* CDI summary means presented in Table 1 indicate that non-clinical children who were not exposed to the traumatic event had significantly lower CDI scores than both clinically-referred non-trauma-exposed children, and non-clinically referred trauma-exposed children. Although it is important to present and discuss results in this way, further between-group examination of CDI item means may yield additional information that is not inferred through results with summary scores. For instance, Table 1 indicates four items in which item means were significantly higher for the non-clinical trauma-exposed children than for the clinical non-trauma-exposed children ("I am sad all the time", "I hate myself", "I want to kill myself", and "I feel like crying everyday"). On the other hand, these hypothetical means indicate three items that clinically-referred children endorsed significantly more frequently than trauma-exposed children ("I am bad all the time", "I never do what I am told", "I get into fights all the time"). Because these three latter items appear to have overlapping features with constructs that are characterized by disruptive behavior, these hypothetical item analysis results appear to have offered greater clarity to results revealed by between-group comparisons with summary scores.

Table 1
Hypothetical CDI item means by group

CDI item	Group differences				
	(1) Trauma-exposed (n = xx)	(2) Clinical non-exposed (n = xx)			
I am sad all the time	0.74 (0.84)	0.45 (0.84)	(3) Non-clinical non-exposed (n = xx)	0.19 (0.57)	1 > 2 > 3
Nothing will ever work out for me	0.47 (0.68)	0.44 (0.68)		0.36 (0.54)	None
I do everything wrong	0.34 (0.61)	0.42 (0.71)		0.19 (0.52)	None
Nothing is fun at all	0.63 (0.69)	0.45 (0.69)		0.37 (0.54)	1 > 3
I am bad all the time	0.13 (0.38)	0.37 (0.58)		0.07 (0.35)	2 > 1 & 3
Terrible things will happen to me	0.72 (0.80)	0.67 (0.80)		0.46 (0.71)	None
I hate myself	0.29 (0.65)	0.09 (0.65)		0.10 (0.38)	1 > 2 & 3
I want to kill myself	0.65 (0.62)	0.41 (0.62)		0.21 (0.45)	1 > 2 > 3
I feel like crying every day	0.40 (0.76)	0.14 (0.76)		0.10 (0.41)	1 > 2 & 3
Things bother me all the time	0.45 (0.77)	0.35 (0.77)		0.18 (0.56)	1 > 3
I do not want to be with people at all	0.40 (0.71)	0.23 (0.71)		0.07 (0.25)	1 > 3
I cannot make up my mind about things	0.68 (0.78)	0.53 (0.78)		0.57 (0.71)	None
I look ugly	0.35 (0.68)	0.25 (0.68)		0.22 (0.51)	None
I have to push myself to do schoolwork	0.51 (0.77)	0.55 (0.77)		0.33 (0.67)	None
Most days I do not feel like eating	0.34 (0.70)	0.19 (0.80)		0.29 (0.68)	None
I worry about aches & pains all the time	0.79 (0.80)	0.57 (0.83)		0.42 (0.69)	1 > 3
I feel alone all the time	0.66 (0.79)	0.45 (0.79)		0.18 (0.42)	1 & 2 > 3
I never have fun at school	0.50 (0.62)	0.47 (0.62)		0.38 (0.64)	None
I do not have any friends	0.53 (0.60)	0.36 (0.55)		0.36 (0.54)	None
I do badly in classes I used to be good in	0.38 (0.69)	0.58 (0.69)		0.42 (0.67)	None
I never do what I am told	0.30 (0.63)	0.61 (0.63)		0.29 (0.56)	2 > 1 & 3
I get into fights all the time	0.24 (0.55)	0.57 (0.69)		0.20 (0.53)	2 > 1 & 3
CDI total	12.23 (8.1)	11.38 (7.7)		7.48 (7.1)	1 & 2 > 3

Note: For each item, the two-point response is listed. Each item entailed a 3-choice response format from 0 (absence of "symptom") to 2 ("symptom" present during the past two weeks).

5. Limitations of item analysis

Despite the potential advantages of item analysis, certain factors may deter researchers from using this procedure as a supplement to summary scores. One important limitation of item analysis is the concern that researchers, when conducting a large number of analyses, may discover an inflated number of statistically significant findings through random chance. That is, significant relations may be reported that in fact represent erroneous findings (i.e., Type I error). Researchers, however, may take steps to decrease the likelihood of Type 1 errors by reducing alpha levels when conducting item analyses (e.g., using $\alpha = 0.01$ or $\alpha = 0.001$). Thus, the statistical concerns that arise from incorporating item analyses can be controlled.

A second potential limitation is that the length of manuscripts may increase when researchers incorporate results obtained via item analysis. Due to space limitations of journals, editors typically regulate the length of manuscripts accepted for publication. One method to reduce the amount of space devoted to item analysis is to use tables (such as the example depicted in Table 1). The author can then refer to the table when describing item analysis results without devoting additional space to specific descriptive and statistical values. A presentation of item analysis results in table format will enable readers to form more complete conclusions based on research results and may facilitate generation of hypotheses and prediction of item relations in future studies.

6. Summary

Behavioral classification provides psychologists with useful ways to communicate empirical findings. However, summary scores and labels (typically derived from indirect methods of assessment) often fail to elucidate behavioral relations sufficiently. As a supplement to analyses involving summary scores and labels, item analyses provide a level of specificity that may enable psychologists to more precisely interpret and predict relations among variables. We therefore encourage researchers to supplement summary-score analyses with item analyses, thus capitalizing on the strengths of indirect methods of assessment, facilitating data interpretation, and offering improved precision and a greater amount of information that may be used to guide theory development.

Acknowledgements

We would like to thank C. W. Lejuez for his helpful comments on earlier drafts of this paper.

References

- Adams, H. E., & Cassidy, J. F. (1993). The classification of abnormal behavior: An overview. In P. B. Sutker, & H. E. Adams (Eds.), *Comprehensive handbook of psychopathology* (2nd ed.). New York: Plenum Press.

- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Brown, T. A., DiNardo, P. A., & Barlow, D. H. (1994). *Anxiety disorders interview schedule for DSM-IV*. New York: Harcourt Brace.
- Cone, J. D. (1978). The Behavioral Assessment Grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy*, 9, 883–888.
- Kaufman, J., Birmaher, B., Brent, D., Rao, U., Flynn, C., Moreci, P., Williamson, D., & Ryan, N. (1997). Schedule for affective disorders and schizophrenia for school-age children – present and lifetime version (K-SADS-PL): initial reliability and validity data. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36, 980–988.
- Kovacs, M. (1992). *Children's depression inventory manual*. North Tonawanda, NY: Multi-Health Systems, Inc.
- Scotti, J. R., Morris, T. L., McNeil, C. B., & Hawkins, R. P. (1996). DSM-IV and disorders of childhood and adolescence: Can structural criteria be functional? *Journal of Consulting and Clinical Psychology*, 64, 1177–1191.
- Sprock, J., & Blashfield, R. K. (1983). Classification and nosology. In M. Hersen, A. E. Kazdin, & A. S. Bellack (Eds.), *The clinical psychology handbook* (pp. 289–307). New York: Pergamon Press.