Rethinking the Use of the Hogan Empathy Scale: A Critical Psychometric Analysis

Robin D. Froman, Suzanne M. Peloquin

Key Words: education, occupational therapy, evaluation process, occupational therapy

Objective. The purpose of this research was to study the stability, internal consistency, factor structure, and convergent and discriminate validity of the Hogan Empathy Scale (EM) when used longitudinally with occupational therapy students.

Method. More than 300 occupational therapy students completed the EM once; 192 completed it twice over a 12-month interval; and 56 completed a third administration at intervals ranging from 3 years to 6 years. The Fieldwork Performance Evaluation (FWPE) was rated twice for students after fieldwork rotations in the occupational therapy program. Data on grade-point average, gender, and age were collected.

Results. Stability was estimated at .41 over a 12-month interval and from .30 to .38 over longer intervals. Internal consistency was estimated at .57, and factor structures hypothesized previously were not replicable. Students’ biographical variables explained only trivial amounts of variance in EM scores in regression equations (R = .08 and .21). Correlations between EM and FWPE scores did not support convergent validity (r = -.01–.18).

Conclusion. The reliability estimates for the EM as a measure of a trait-like construct are not encouraging and do not replicate previous estimates. Validity evidence was equally disappointing, raising questions about what the EM is measuring and cautioning against its continued, uncritical use as a measure of empathy.


Much has been written over the years and across a number of disciplines on the topic of empathy, with peak interest occurring during the 1970s and early 1980s. A resurgence of research on empathy has appeared within the past decade, probably as a result of the increasing focus on efforts to humanize health care. Those disciplines contributing to the literature on empathy include occupational therapy and other allied health professions, nursing, counseling, developmental psychology, education, medicine, the medical humanities, neurology, psychiatry, psychology, and social work.

The American Occupational Therapy Association (AOTA, 1993) cited the construct of empathy as a requisite attitude for practice. This end-of-the-century article reflects values found in guiding documents since the beginning of the century. For example, Tracy, a nurse invited to the profession’s founding session, described her empathy in contemporary terms: “The value of wise human sympathy, of cheerfulness in work and mien, of tactful dealing with unreasonableness and irritability, of skillful diversion of...
thought from pessimistic channels, are essential parts of the nurse's equipment to do her work" (Tracy, 1913, pp. 9–10).

Kidner (1929) expressed support for similar attitudes: "Kindness, humanity, decency, honor, and good faith—to give these up under any circumstances whatever would be a loss greater than any defeat, or even death itself" (p. 385). World War I reconstruction aide Ruggles captured the same view metaphorically: “You must reach for the heart as well as the hands. It’s the heart that really does the healing” (Carlota & Ruggles, 1946, pp. 249–250).

In the 1980s, health care leaders reiterated the commitment to shaping a climate of caring within health care systems increasingly associated with impersonal behaviors (Baum, 1980; King, 1980; Yerxa, 1980). Mentioned as central to this effort was empathy. Mosey (1981) identified empathy as a capacity central to the art of practice, a necessary complement to its science. Not surprisingly, occupational therapists devoted attention to the nature and development of empathy in practice (Christiansen, 1977; Peloquin, 1989, 1995; Wise & Page, 1980) and to the shaping of interpersonal behaviors that derive from empathy (Delworth, 1972; Furnham, King, & Pendleton, 1980; Lloyd & Maas, 1991; Peloquin & Davidson, 1993; Tickle-Degnen, 1998). Others in occupational therapy have considered and commented on constructs associated with empathy: interactive reasoning (Fleming, 1991), intersubjective understanding (Crepeau, 1991), client-centered service (AOTA, 1995), collaboration (Bowen, 1996), and good occupational therapy practice (Wright-St. Clair, 2000).

The occupational therapy profession is moving forward in its commitment to graduate-level education and research into the nature of effective interventions. The knowledge, skills, and attitudes that lead to positive outcomes will drive future practice. One question the profession is contemplating is: What factors contribute to effective partnerships between consumers and practitioners that foster and enhance participation of persons with or at risk for disabling conditions? (American Occupational Therapy Foundation, 2000). Answers to such questions will turn on the use of tools thought rigorous enough to measure empathy and other requisite attitudes among students and practitioners in the field.

Unfortunately, many studies proposing to investigate empathy and its relation to health education, health care, and health care practitioners relied on and continue to rely on instruments developed decades ago. Use of dated instruments can interfere with answering research questions regarding the relationships among empathy, health care education, and health care delivery. This instrumentation study was part of a larger, longitudinal study of empathy in relationship to other constructs in occupational therapy students. Because earlier studies in occupational therapy had used the easily administered, self-report Hogan Empathy Scale (EM; Hogan 1969) and because the scale was so widely cited in current health care literature, the EM was selected for use as a measure of the cognitive aspect of empathy. Unfortunately, initial inspection of data on the performance of the EM raised questions about its suitability for use in the project. In particular, the following questions emerged regarding the EM and directed this study:

- Do current reliability estimates (stability and internal consistency) for the EM replicate those reported previously and those expected for a trait-like measure of empathy?
- Is the EM a unidimensional or multidimensional measure, and does the factor structure of the EM correspond to that initially expected and reported for the scale?
- Is there current, supportive evidence for the convergent and discriminative validity of EM scores?

The Conceptualization and Development of the Hogan Empathy Scale

Psychologist Robert Hogan reported the development of the EM in 1969. He set out to examine the contemporary dictionary definition of empathy as “the intellectual and imaginative apprehension of another's state or condition without actually experiencing that person's feelings” (Hogan, 1969, p. 308) and to determine whether this social phenomenon was recognized among lay persons as well as psychologists. Hogan likened his orientation to that of role theorists Cottrell (1971), Gough (1965), Mead (1934), and Sarbin and Allen (1968), who saw empathy as taking the role of the generalized other. Hogan (1969) agreed with Mead that empathy was the essence of social intelligence; he further believed that empathy related to moral development as a “disposition to adopt a broad moral perspective” (p. 307) and that it was trait-like in nature. Being trait-like, by definition, implies the construct should show stable measurement qualities over time.

Hogan (1969) asked teachers, clerks, and psychology students (N = 23) to select 50 items from the 100-item California Q-sort descriptors to characterize “a highly empathic man” (p. 308). Correlations among the items, when responded to for ratings of real people, suggested to Hogan that empathy was a relatively discrete and recognizable phenomenon. It also appeared that the construct was unidimensional (Block, 1961). Two groups of study participants were subsequently rated on the proposed items for the empathy scale. The first group consisted of 100 military officers, the second 45 research scientists and student engineers. The study participants were categorized into high and low empathy groups according to their ratings on an overall empathy criterion. Their responses were compared on 957 true–false items in the California Psychological Inventory (CPI), the Minnesota Multiphasic Personality Inventory, and an Institute of Personality and Research pool of items. Statistical analyses (chi-square, Fisher's exact test) were used to select the most discriminating items to construct the final 64-item EM.
Initial and Subsequent Reliability and Validity Estimates for the EM

Hogan has reported extensive analyses of reliability and validity of the 64-item scale conducted among 11 groups of men and 3 groups of women (N = 1,086) (Hogan, 1969). He noted that the limited study of women showed them to score slightly higher on the scale, on average, than men, but he provided no conceptual basis for this difference. Hogan concluded that the EM had adequate reliability (the lowest test–retest correlation after 2 months was .71 among 50 college undergraduates) and supportive evidence of concurrent validity (i.e., correlations with scales such as the Social Acuity scale of the CPI). Greif and Hogan (1973) again reported satisfactory psychometric properties: a test–retest reliability coefficient of .84 over a 3-month interval and an internal consistency estimate of .71. They discussed “conceptual validity” as the extent to which the scale corresponded with the conceptual framework within which it was developed. They concluded a “reasonable correspondence” after the reporting of several studies, such as one in which the EM successfully discriminated between two samples (130 college students and a sample of 119 incarcerated delinquents) in a known group. Other investigators compared EM scores with other specific indicators of empathy. Christiansen (1977) found that EM scores correlated with peer ratings and faculty ratings of empathy in occupational therapy students (range = .39–.57). Hornblow, Kidson, and Jones (1977) found that EM scores for medical students related to empathy ratings given by students’ peers (r = .45) but were unrelated to course grades in psychiatry, self-ratings of empathy, and patient ratings on a Relationship Inventory of empathy. A summary of these studies of the EM when used with health professions students suggests that concurrent validity evidence is equivocal.

Use of the EM Among Professional Students

Deardoiff, Kendall, Fich, and Sitarz (1975) administered the EM, Spielberger’s State–Trait Anxiety Inventory, and Rotter’s Locus of Control Scale to 59 college students. Empathy scores correlated significantly with both state and trait anxiety but were independent of locus of control, suggesting that each scale tapped unique interpersonal dimensions. Kupfer, Drew, Curtis, and Rubinstein (1978) reported no relationship between the short version (47-item) of the EM and Medical College Admissions Test scores and no correlation with variables thought to lead to clinical success for medical students. Bussa (1993) later reported no evidence of expected relationships between eight biographical variables and empathy measured by the EM in a sample of 210 nursing students.

Other investigators compared EM scores with other specific indicators of empathy. Christiansen (1977) found that EM scores correlated with peer ratings and faculty ratings of empathy in occupational therapy students (range = .39–.57). Hornblow, Kidson, and Jones (1977) found that EM scores for medical students related to empathy ratings given by students’ peers (r = .45) but were unrelated to course grades in psychiatry, self-ratings of empathy, and patient ratings on a Relationship Inventory of empathy. A summary of these studies of the EM when used with health professions students suggests that concurrent validity evidence is equivocal.

Other studies, however, suggest the EM may be an adequate indicator of empathy. Holt-Ashley (1985) evaluated an empathy training intervention by giving the EM to 50 nursing students and discovered an inverse relationship between age and empathy level. EM scores in the experimental group (n = 24) increased after empathy training classes; EM scores of the control group (n = 26) decreased from pretest to posttest, suggesting EM sensitivity to treatment effects.

In summary, the results of studies that correlated EM scores and a variety of other measures are inconsistent, and a strong consistent pattern of replication of results across studies does not exist. Findings of significant differences in the intervention study with nursing students is interesting (Holt-Ashley, 1985). The mixed results of the total body of literature on the EM suggest the need for more research with attention to psychometric estimates of the scale’s reliability and validity. The purpose of this study was to identify the psychometric estimates relevant for the EM for current use.
Method

Participants

From 1991 to 1999, occupational therapy students in an undergraduate program in the United States participated in a research project aimed at developing a better understanding of the characteristics of occupational therapy students. The EM was one of the measures used in that project. The university Institutional Review Board reviewed and approved the project. Participants used a code number that guaranteed their anonymity while allowing matching of cases over time and facilitating access to records for data collection by support staff blind to the purpose of the study. Participants completed a demographic sheet and the 64-item EM on the initial round of data collection during their junior year of study. They completed the EM again during their senior year. Academic records were accessed for scale scores and subscores on the Fieldwork Performance Evaluation (FWPE; AOTA, 1986) and for entering and exiting grade-point averages (GPAs). GPAs for 1992 were not available because of differences in clerical management that year. The FWPE was completed at the end of clinical placements, for a total of two evaluations per participant during the occupational therapy academic program. In 1997, the EM was sent to participants, then alumni, for whom mailing addresses were available. The interval between first and third administration of the EM varied from 3 to 7 years, depending on the year of graduation. The interval between second and third administration varied from 2 to 6 years.

The FWPE

In the larger research project from which these instrument validation data were derived, the FWPE was used to assess a student's acquisition of the requisite skills for occupational therapy practice. Scores from the FWPE were secured from student records because evaluation on the instrument was part of the clinical portion of the occupational therapy educational program. The FWPE is composed of three major scales: Performance, Judgment, and Attitude. Items are rated on 5-point rating scales (1 = poor, 5 = excellent) by fieldwork supervisors (i.e., faculty members or practitioners observing students). The FWPE was evaluated with 597 field test forms completed and submitted for data analysis (AOTA, 1986). Reliability estimates for the three major scales show high stability and homogeneity (Cronbach's coefficient alphas), with estimates falling between .92 and .96 for the Performance, Judgment, and Attitude scales. These major scales have five subscores for the clinical functions of treatment, assessment, planning, problem solving, and administration and professionalism.

Two FWPE ratings are of particular interest in studies related to interpersonal effectiveness and empathy: the attitude scale and the treatment subscore within the Performance scale. The Attitude scale pertains to affective expressions and perceptions. The treatment subscore reflects skill at interpersonal behavior and interpersonal relationships when actually interacting with others and, thus, would be most closely conceptually aligned with empathy as defined by Hogan (1969). Given these definitions of the scale and subscore ratings, these specific measures on the FWPE were expected to show the highest relationship, as convergent validity estimates, with EM scores.

Data Analysis

A total of 320 participants completed the study instruments at least once. Smaller numbers completed the EM scale twice (n = 192), and a subgroup completed the EM scale at all three data collection points (n = 56). The majority of the biographical characteristics are similar across data sets with one, two, or three administrations available for the EM scale. Data on marital status were collected to describe the sample and not used in further analyses. Year of graduation was used to calculate the interval between second and third testing and to examine whether rough equivalence of representation of graduating classes existed in the sample. The biographical data for the total group and for each subgroup of participants are presented in Table 1.

Data were analyzed in three steps corresponding to the research questions posed. First, correlational analyses were conducted to assess reliability. Both pairwise correlations for stability estimation and Cronbach's alpha estimate for internal consistency were calculated.

The second research question, pertaining to the dimensionality of the EM, was addressed through the use of factor analysis procedures. Finally, the third research question regarding convergent and discriminate validity evidence was approached through the use of correlational analysis.

Results

Reliability

The first research question pertains to the reliability of EM scores. First, pairwise correlations were calculated between EM scores at junior year, senior year, and postgraduation follow-up (2 to 6 years after graduation). These correlations represent the relationship between two measures of the same construct, empathy, over time. As such, they are estimates of the stability of the EM scores. As reliability estimates, the correlations may be interpreted as direct estimates of the shared variation between the scores at different times (DeVellis, 1991; Nunnally & Bernstein, 1994). The correlation of junior year and senior year EM scores was .41; the junior year to postgraduation follow-up correlation was .38; and the senior year to postgraduation follow-up correlation was .30. The junior year to senior year correlation, indicating stability of scores while in the occupational therapy educational program, shows about
60% of the variation of scores at the end of the program to be independent of EM scores as measured at the junior year. Similarly, the senior year correlation with postgraduation follow-up suggests that about 70% of the variation in EM scores is independent of scores measured at the end of the program. The correlation between junior year and the postgraduation follow-up scores, the stability estimate over the longest interval, shows about 62% independence of EM scores at the two most distant measurement points. Although all correlations were significant (p < .05), they were modest, at best, as stability estimates for a construct conceptualized as trait-like.

The second reliability estimate, that of internal consistency for the EM, was calculated with Cronbach’s alpha. When calculated on the total sample (N = 320), the estimate was .57 for the entire 64-item scale, which is notably less than the alpha estimates of .71 (Greif & Hogan, 1973) and .69 (Johnson et al., 1983) reported previously. The estimate shows relatively low internal consistency for such a long scale and is substandard for a reliability estimate for an established instrument (Nunnally & Bernstein, 1994).

Table 1
Biographical Description of the Total Sample and Subsamples

<table>
<thead>
<tr>
<th>Variable</th>
<th>Completed Total Sample</th>
<th>Completed Questionnaires Twice</th>
<th>Completed Questionnaires Thrice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (M/SD)</td>
<td>26.8 (7.22)</td>
<td>27.1 (7.25)</td>
<td>26.5 (7.51)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>43 (13%)</td>
<td>24 (12%)</td>
<td>3 (5%)</td>
</tr>
<tr>
<td>Female</td>
<td>189 (59%)</td>
<td>111 (58%)</td>
<td>39 (70%)</td>
</tr>
<tr>
<td>Unidentified</td>
<td>88 (28%)</td>
<td>57 (30%)</td>
<td>14 (25%)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>145 (61%)</td>
<td>112 (58%)</td>
<td>32 (57%)</td>
</tr>
<tr>
<td>Married</td>
<td>77 (32%)</td>
<td>62 (32%)</td>
<td>17 (30%)</td>
</tr>
<tr>
<td>Divorced</td>
<td>14 (6%)</td>
<td>14 (7%)</td>
<td>6 (11%)</td>
</tr>
<tr>
<td>Widowed</td>
<td>2 (1%)</td>
<td>1 (&lt; 1%)</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Widowed</td>
<td>1 (&lt; 1%)</td>
<td>1 (&lt; 1%)</td>
<td></td>
</tr>
<tr>
<td>Unidentified</td>
<td>—</td>
<td>2 (1%)</td>
<td></td>
</tr>
<tr>
<td>Year graduated</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>39 (12%)</td>
<td>32 (17%)</td>
<td>13 (23%)</td>
</tr>
<tr>
<td>1993</td>
<td>37 (12%)</td>
<td>15 (8%)</td>
<td>6 (11%)</td>
</tr>
<tr>
<td>1994</td>
<td>43 (13%)</td>
<td>21 (11%)</td>
<td>8 (14%)</td>
</tr>
<tr>
<td>1995</td>
<td>42 (13%)</td>
<td>26 (14%)</td>
<td>14 (25%)</td>
</tr>
<tr>
<td>1996</td>
<td>42 (13%)</td>
<td>32 (17%)</td>
<td>15 (27%)</td>
</tr>
<tr>
<td>1997</td>
<td>41 (13%)</td>
<td>30 (16%)</td>
<td>15 (27%)</td>
</tr>
<tr>
<td>1998</td>
<td>39 (12%)</td>
<td>13 (7%)</td>
<td>15 (27%)</td>
</tr>
<tr>
<td>1999</td>
<td>37 (12%)</td>
<td>23 (12%)</td>
<td></td>
</tr>
<tr>
<td>GPA (M/SD)</td>
<td>3.52 (0.28)</td>
<td>3.57 (0.25)</td>
<td>3.56 (0.24)</td>
</tr>
</tbody>
</table>

Note. GPA = grade-point average. All percentages rounded to nearest whole number; some sums do not equal 100% due to rounding.

Validity

The third research question was directed at the study of the convergent and discriminate validity evidence for the EM. Specifically, the relationship between variables conceptually similar to empathy (convergent validity) and those that should be independent of empathy given Hogan’s conceptualization (discriminate validity) were explored. The scores from the FWPE intended to measure interpersonal skills, relationships, and psychosocial attitudes similar to empathy were expected to show convergent validity evidence with EM scores and to correlate in a statistically and practically meaningful way. Given Hogan’s definition of empathy as a cognitively based construct reflecting the ability to apprehend another’s mental state, biographical variables such as...
Age, gender, and GPA were expected to be independent of EM scores.

The FWPE score on the Attitude scale and the treatment subscore of the Performance scale were expected to show convergent validity with EM scores. Two pairs of correlations (one calculated from data collected from the junior year, the second calculated from subsequent data collected from the senior year) were available for FWPE and EM scores. The concurrent correlations of the Attitude scale were low, showing correlations of .19 and .07 with EM. For the treatment subscore of the Performance scale, the correlations with EM scores were .16 and .01. As concurrent convergent validity estimates, or two simultaneous measures of constructs expected to be related, these correlations, when squared, may be interpreted as the percentage of shared or common variation between the measures. Hence, between 0% and 3.6% of the variation in EM and FWPE subscale scores is common variation. Thus, although the correlations of .16 and .19 are significant, they represent very little common variation between EM and FWPE subscale scores expected to correspond to empathy.

The second part of the validity question, that related to discriminate validity, was studied through inspection of correlations of EM scores and biographical variables of age, gender, and GPA. Pairwise correlations of each variable with EM scores, at both junior and senior year rounds of data collection, are presented in Table 2. The correlations range from .01 to .13 and represent between 0% and 1.7% shared variation between EM scores and participants’ biographical characteristics. These correlations clearly indicate that empathy, as measured with the EM, is independent of age, gender, and GPA. These data support EM scores as a measure of empathy being independent of other personal characteristics that conceptually should be unrelated to empathy.

Discussion
The intent of this validation study of the EM was to update the psychometric estimates available for the scale, analyze critically the scale’s properties of reliability and validity, and assess whether scale properties align with current expectations and conceptualizations of empathy. The reliability analyses, looking specifically at the stability and internal consistency estimates for the EM, were disappointing. The stability of the EM scores showed considerable variation over a 1-year or longer interval and is not a supportive finding for a measure of a construct conceptualized as a trait. The EM total scale internal consistency estimate of .57 using a Cronbach’s alpha formula for calculation is considerably less than the .80 generally accepted as the recommended value for established scales (Nunnally & Bernstein, 1994).

The results of the factor analyses conducted on participants’ responses supported neither a single-factor, a three-factor, nor a four-factor structure for responses. A single-factor solution is implied by the recommendation to score the EM with a single total scale score. Both the three-factor and four-factor structures had been reported in the literature previously. The lack of support for any of these factor solutions in the present data is disappointing and not consistent with expectations about EM performance. Additionally, the findings do not align with more current models of empathy as a multidimensional construct (Williams, 1990). In short, the factor structure of the EM is elusive and incongruent with the original literature of the time when it was developed, subsequent study of the EM, and current observations about the construct of empathy.

Finally, only the convergent and divergent validity study offered slight supportive psychometric evidence for the EM. The discriminate validity analysis showed EM scores in this study to be independent of age, gender, and GPA. This finding was expected because those personal characteristics would not be expected to be related to empathic ability as conceptualized by Hogan. Unfortunately, the convergent validity analyses did not support the relationship between EM scores and FWPE scores most conceptually associated with empathy. EM and FWPE scores, rated two separate times by independent faculty members or clinical fieldwork supervisors, did not show practically meaningful relationships with EM scores.

Conclusion
The results of the validation analyses reported here calls into question the continued, uncritical use of the Hogan Empathy Scale to study empathy and its correlates in research in the health professions. The basic foundational psychometric estimates of internal consistency of the EM fall well below those recommended for instruments used in research studies. The lack of even modest stability raises questions about conceptualizing the construct as a trait when measured with the EM. In the absence of acceptable reliability estimates, it is not surprising that the subsequent factorial, convergent, and discriminate validity analyses were disappointing. Given these findings, researchers may wish to suspend their use of the EM. The present study has not proposed to either alter the scale or offer an alternate scale for use. Rather, the findings of our validation analyses are shared here to caution others before they employ the EM in a lengthy, time-consuming project that might yield disappointing findings as a result of poor measurement. ▲
Acknowledgment
We thank Kim Vogel, EdD, OTR, for assistance in securing the Hogan Empathy Scale.

References