Empathy Training: Methods, Evaluation Practices, and Validity

Tony Chiu Ming Lam
Klodiana Kolomitro
Flanny C. Alamparambil
University of Toronto

Background: Empathy is an individual's capacity to understand the behavior of others, to experience their feelings, and to express that understanding to them. Empathic ability is an asset professionally for individuals, such as teachers, physicians and social workers, who work with people. Being empathetic is also critical to our being able to live with others in general, and ultimately to leading happier lives. Subsequently it seems imperative to examine if and how it is possible to enhance people's empathic ability.

Purpose: The purpose of this article is to use narrative review method to analyze studies of empathy training in human service and social science disciplines over the past thirty years to address the questions: "How have people been trained in empathy and what are the findings?" and "How was empathy training evaluated and how valid are these evaluation findings?"

Setting: Not applicable.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Narrative review.

Findings: Twenty-nine articles pertaining to empathy training evaluation research were identified based on an advanced computer search on the following databases: "Education Full Text," "ProQuest Education Journals," "Web of Knowledge" and "Educational Resources Information Center (ERIC). Seven types of training methods were noted in these 29 evaluations with the most popular being didactic related (42%). All but two studies (93%) reported positive findings, mainly in regard to learning (86%), or the cognitive component of empathy. These findings suggest that regardless of the training method, individuals can learn about the concept of empathy. Unfortunately, information pertaining to the effects of training on individuals' feeling for others, and their ability and propensity to take the perspective of others and to demonstrate it in the natural environments is lacking. Consequently, very little is known about the trainability of the affective and behavioral components of empathy. Also, some of the findings were moderated by gender, age, education level, and time of measurement. Regarding evaluation research designs, most of the studies used self-reporting to collect trainees' knowledge about empathy and most of the quantitative studies used a control group and pretesting to examine training impact. Construct validity of both empathy measurement and training is very problematic. A majority of the studies did not clearly define empathy, provide training as defined, and/or measure what is being trained; conceptualization of empathy across studies was not consistent either. In sum, data from the studies reviewed were neither complete nor valid enough to provide a clear and full understanding of the trainability of empathy. More research is apparently needed and hopefully lessons learned from our review will be considered in designing future studies.

Keywords: *empathy training; empathy training evaluation; narrative review*

_____

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C. Alamparambil

To address the questions: "How have people been trained in empathy and what are the findings?" and "How was empathy training evaluated and how valid are these evaluation findings?" we used the narrative review method to analyze 26 quantitative and three qualitative studies of empathy training in human service and social science disciplines over the past thirty years. Seven types of training methods were noted in these evaluations with the most popular being didactic related (42%). All but two studies (93%) reported positive findings, mainly in regard to learning (86%), or the cognitive component of empathy. These findings suggest that regardless of the training method, individuals can learn about the concept of empathy. Unfortunately, information pertaining to the effects of training on individuals' feeling for others, and their ability and propensity to take the perspective of others and to demonstrate it in the natural environments is lacking. Consequently, we know very little about the trainability of the affective and behavioral components of empathy. Also, some of the findings were moderated by gender, age, education level, and time of measurement. Regarding evaluation research designs, most of the studies used self-reporting to collect trainees' knowledge about empathy and most of the quantitative studies used a control group and pretesting to examine training impact. Construct validity of both empathy measurement and training is very problematic. A majority of the studies did not clearly define empathy, provide training as defined, and/or measure what is being trained; conceptualization of empathy across studies was not consistent either. In sum, data from the studies reviewed were neither complete nor valid enough to provide a clear and full understanding of the trainability of empathy. More research is apparently needed and hopefully lessons learned from our review will be considered in designing future studies.

Empathy is an individual's capacity to understand the behavior of others, to experience their feelings, and to express that understanding to them. Subsequently, there are three components of empathy: cognitive, affective, and behavioral. The cognitive component refers to one's ability to take the perspective of others and see the world through their perspective. Dymond (1949) refers to this aspect of empathy as the "imaginative transposing of oneself into the thinking, feeling and acting of another and so structuring the world as he does" (p.127). It must be noted that the cognitive aspect of empathy does not refer to intellectually knowing the concept of empathy, but rather having the ability to take others' perspectives. The affective component of empathy involves experiencing the feelings of another person. Barrett-Lennard (1981) proposed a theory that empathy must involve "resonating" with another person's emotions (in Kagan & Schneider, 1987, p. 459), where a person "physiologically experiences the other person's affects" (Holm, 1996, p. 241). The behavioral component involves verbal and non-verbal communication to indicate an understanding of an emotional resonance with the other person. Kagan and Schneider (1987) specify that empathy requires "a person to communicate...that he or she has perceived another person's message" (p. 460).

Although generally we perceive individuals as empathetic if they are able to act genuinely in the best interests of others (that is, possess all three empathy components), the literature is not clear whether individuals can possess only one

or a combination of the three empathy components, and if they would be considered empathetic if they did not possess all three abilities. Kagan and Schneider (1987) propose that the three components of empathy not only exist, but all are present in an empathetic person. In fact, in their conceptualization, empathy occurs in phases such that a person first experiences emotional empathy, then cognitive empathy, then displays behavioral empathy. While the sequence of occurrence is conjectural, there is empirical evidence that at least some of these components of empathy are positively correlated. For example, significant correlation has been reported between responses to emotional and cognitive subscales of the Interpersonal Reactivity Index (Davis, 1980). Davis points out that the correlations are not so high as to lead us to conclude that the subscales are measuring the same construct, but that they must be different aspects of the same construct. Jackson (1985, in Kagan & Schneider, 1987) reported statistically significant relationships between participants' scores on the Affective Sensitivity Scale and peer evaluation of some desirable behavioral manifestations of empathy such as warmth and openness.

Empathic ability is an asset both personally and professionally. Empathic expression with a partner enhances relationship satisfaction (Long et al. 1999; Ridley et al. 1982)**.** As well, empathy should augment the ability of professionals who work in human services to bring about change in the people they are helping. For example, empathetic teachers should be more effective in inspiring students to change poor work habits and to learn than non-empathetic teachers, because they are able to view the world from the students' perspectives and hence more likely to connect with them. Findings from Aspy, Roebuck, and Aspy (1984) research appear to support that hypothesis as they found that teacher empathy was positively correlated with student attendance and student achievement on tests. Herbek and Yammarino (1990) cite several studies showing empathy as an important factor in the success of several professions. They wrote:

> Coffman (1981) and Bochner and Yerby (1977), among others have demonstrated that empathy is an important instructor variable that positively affects learning outcomes by creating a psychologically safe learning environment. Von Bergen and Shealy (1982) indicate that empathy training for salespeople is a key component in successful selling. Empathy is also a significant part of supervision and one of the core components of the effective supervisor's skills (Boyd, 1978). Moreover, empathy has been identified as key aspect of the practice patterns of successful and litigation-free physicians (Reiser & Rosen, 1984) and has been shown to be essential for effective leadership in groups and organizations (Bas, 1981). (Herbek & Yammarino, 1990, p. 281.)

If being empathetic is so critical to our being able to live and work congenially with others, and ultimately to leading happier lives, it seems logical to ask the question: Is it possible to enhance people's empathic ability?

In our initial review of the literature, it appears that research has shown that training can enhance one's empathy. For example, eight out of nine empathy training studies that Layton (1979) cited were successful in inducing empathic behavior through modeling. Long, Angera, and Hakoyama (2006) found an increase in empathy between husband and wife when videotapes of the couple's

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C. Alamparambil

argument were played back to them. Barone, Hutchings, Kimmel, Traub, Cooper, and Marshall (2005) were able to train students in graduate psychology courses to become more empathetic. The collective findings from the research suggest that empathy can be enhanced through training. However, there are also findings that raise questions about both the trainability of empathy and the generalizability of the observed training effects. For example, Beddoe and Murphy (2004) reported no increase in empathy of nurses after eight weeks of training. As well, Seto, Young, Becker, and Kiselica (2006) reported no increase in empathy in counselors-in-training after a six-week training program. Barone et al. (2005), Fernandez-Olano, Montoya-Fernández, and Salinas-Sánchez (2008), Avery and Thiessen (1982), and Long, Angera, Carter, Nakamoto, and Kalso (1999) noted differential training effects for male and female trainees, thus casting doubt about the extent to which empathy training findings can be generalized across different populations. In addition, it appears that researchers conceptualize and measure empathy as well as training impact differently. Also, methodologies used in some of these studies are questionable, consequently casting doubt on the validity of some research findings.

In order to further our understanding and to provide suggestions and direction for future research on the trainability of empathy, it seems imperative that researchers go beyond a general literature review. It is important to systematically delve into the literature to analyze and synthesize findings from empathy training studies and to identify areas of improvement for future studies in order to enhance validity of new findings. This is the goal of our research. We conducted a narrative review (Shaddish, Cook, &

Campbell, 2002) guided by the following two broad questions:

- How have people been trained in empathy and what are the findings?
- How was empathy training evaluated and how valid are these evaluation findings?

In this paper, we report and discuss our findings. First, we present evaluation findings pertaining to the various methods used to train empathy, and then we discuss the validity of these findings. Finally, we draw our conclusions and offer our opinion regarding future directions and research on the trainability of empathy.

Our literature search strategy involved several actions. We conducted a computer search on the following databases "Education Full Text," "ProQuest Education Journals," "Web of Knowledge" and "Educational Resources Information Center (ERIC)". Search results yielded a combination of electronic journals, and hard-copy printed books and journals. For all searches the advanced search feature was used and two or more word keyword searches were performed (i.e. empathy and train* or training program, empathy and teach* and so on) and no other restrictions were placed on the search, in other words the search was open to utilize all available databases. After we located a few relevant documents, we used the descriptors attached to the record to locate similar records.

We found a total of 29 articles that report results of research studies that evaluate the effectiveness of an empathy training program. (Studies that train children with special needs were excluded because the training was aimed at helping these individuals become more responsive

rather than more empathetic.) A summary of key features of the studies reported in the 29 articles we identified is presented in Table 1.

As seen in Figure 1, the 29 studies span across various groups such as education (24%), nursing (14%), therapy (7%), medicine (21%), social work (3%), psychology (7%), human service (7%), couples (10%) and divorcees (3%). Among the human service disciplines (excluding couples and divorcees), researchers trained students (71%) more often than professionals (one of the 29 studies did not specify their participants). In our further analysis, we did not find that the discipline in which empathy was studied was associated with the type of empathy training or the type of instruments or methods used to measure empathy. Regarding research methodology, 26 (90%) of the studies used quantitative research methods and three (10%) qualitative research methods.

Empathy Trainability Findings by Training Designs

In the 29 studies that we reviewed, we noted that a variety of training methods used in empathy training over the last thirty years. From our analysis of these training methods, we classified them into the following types: (1) *experiential,* (2) *didactic and experiential,* (3) *skill training,* (4) *didactic and skill training,* (5) *mindfulness training,* (6) *video stimulus,* and (7) *writing.* (See Table 1 for the training method used in each study.)

To respond to the first broad question of our paper: "How have people been trained in empathy and what are the findings?" we describe below the seven methods educators used to train empathy

that we identified and the evaluation findings associated with each.

Empathy Training Methods

Experiential Training. As the name implies, *experiential* training emphasizes gaining experience on the part of the trainees to be a critical factor in meaningful learning. In experiential training, the instructors are facilitators who design experiences for trainees. That is all they do; there is no lecturing on theory and concept. The most influential writer on experiential training is Kolb (1984). His model, referred to as the Reflective Learning Cycle, consists of four phases in which the trainer provides an experience and then (1) the participants reflect on it, (2) formulate guiding principles, (3) apply the learning, and (4) receive feedback. This particular method of training was employed in 2 of the 29 studies (7%) we reviewed (Barak, Engle, Katzir, & Fisher, 1987; Feighny, Monaco, & Arnold, 1995).

Barak et al. (1987) provided empirical evidence supporting the effectiveness of experiential training (through the use of a game) in improving the participants' empathy skills. Similarly, preliminary findings of Feighny et al. (1995) suggested that experiential training was effective in increasing the empathy levels of medical students. However, for experiential training approaches that provide multiple experiences, it is not clear which particular experience contributes to empathy enhancement. For example, even though practical experience gained from simulations and games is thought to motivate and help participants develop empathy, it is difficult to determine the

Table 1
Description of Research Designs and Findings of the 29 Studies Reviewed

Study	Discipline	Training Content*	Measure-ment Content*	Training Method	Measurement Instrument	Experimental Design	Sample Size	Significance of Findings	Moderating Variable(s)	Key Findings
Ancel (2006)	Nursing (nurses)	B	A, B, C	Didactic & Skill	Dokmen Scale	Single Group Pre & Post	190	Significant	Age; Education level	Increase in self-examination skills, active listening, understanding others' perspective, and verbal and non-verbal communication practices
Aspy et al. (1984)	Education (teachers)	B	B	Didactic & Experiential	Carkhuff Scale	Control Group Pre & Post	300	Significant	None	Positive relationship between teachers' gain in interpersonal skills and students' classroom performance
Avery & Thiessen (1982)	Divorcees	Unclear	B	Didactic & Experiential	Acceptance of Other Scale	Control Group Pre & Post	13	Significant	Gender; Treatment by gender	Increase in perceived level of social support, responding skills and ability to self-disclose
Barak et al. (1987)	Psychology (students)	A, B, C	C	Experiential	Observer ratings	Randomized Control Group Pre & Post	12	Significant	None	Increase in the ability to sense and understand what others are thinking or feeling and to communicate this understanding
Barone et al. (2005)	Psychology (students)	C	C	Video Stimulus	Observer ratings	Control Group Pre, Post &Follow-up	27	Significant	Gender	Greater accuracy of inferring feelings, but not thoughts.
Beddoe & Murphy (2004)	Nursing (students)	Mindfulness	A, C	Mindfulness	Interpersonal Reactivity Index (IRI)	Single Group Pre &Post	16	Non-significant	None	No change in ability to feel compassion for others and adopt others' perspective; however, participants reported reduced anxiety, greater self-confidence, greater well-being, and ability to be more hopeful and assertive

Study	Discipline	Training Content*	Measure-ment Content*	Training Method	Measurement Instrument	Experimental Design	Sample Size	Significance of Findings	Moderating Variable(s)	Key Findings
Block-Lerner et al. (2007)	Unclear	Mindfulness & B	A, C	Mindfulness	IRI	Single group Post only	40	Significant	None	Greater capacity to take other' perspectives, and ability to write more about other people and less about oneself
Crabb et al. (1983)	Human Service (lay persons in churches)	B, C	B	Didactic & Skill	Empathic Understanding in Interpersonal Processes – A Scale for Measurement (EUIPASM)	Randomized Control Group Post &Follow-up	18	Significant	Group by Condition (Tape)	Increase in communication skills
DasGupta & Charon (2004)	Medicine (students)	A & C	A	Writing	Open-response survey questions	Qualitative Study	11	Irrelevant	None	Explicit awareness of and reflection on personal illness experiences brought participants closer to the experiences of their patients
Evans et al. (1993)	Medicine (students)	Unclear	A, B, C	Didactic & Skill	IRI & Carkhuff Scale	Randomized Control Group Pre & Post	27	Significant	None	Increase in responding skills; no significant differences in IRI's affective and cognitive empathy measures
Feighny et al. (1995)	Medicine (students)	A, B, C	A, B, C	Experiential	IRI & Wolf's Medical Helping Relationship Inventory	Control Group Pre & Post	Unknown	Significant	None	Increase in physician-patient communication skill but not IRI's affective and cognitive empathy measures
Fernandez-Olano et al. (2008)	Medicine (students &residents)	B	A, B, C	Didactic & Skill	Jefferson Scale of Physician Empathy (JSPE)	Control Group Pre & Post	101	Significant	Gender	Increase in perspective taking ability, compassionate care, and ability to "stand in the patients shoes."
Fraser & Vitro (1975)	Education (preservice teachers)	B	B	Video Stimulus	EUIPASM	Control Group Pre & Post	Unknown	Significant	None	Increase in responding skills

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C. Alamparambil

Study	Discipline	Training Content*	Measure-ment Content*	Training Method	Measurement Instrument	Experimental Design	Sample Size	Significance of Findings	Moderating Variable(s)	Key Findings
Gantt et al. (1980)	Human Service (mental health/human service students)	B	B	Didactic & Experiential	Recognition Assessment – Empathy Scale	Single Group Pre, Post &Follow-up	47	Significant	None	Increase after training and maintenance and increase over time in responding skills
Hatcher et al. (1994)	Education (high school &college students)	B	A, C	Didactic & Skill	IRI	Comparison Group Pre & Post	35	Significant	Age	Greater readiness in college but not high school students for learning how be compassionate and take others' perspective
Haynes & Avery (1979)	Education (high school students)	B	B	Skill	Observer ratings & open-responses to vignettes	Randomized Control Group Pre & Post	24	Significant	None	Increase in ability to self-disclose and to respond in different situations with parents, peers, and dating partners
Herbek & Yammarino (1990)	Nursing (nurses)	A, B	A	Skill	Questionnaire Measure of Emotional Empathy	Partial Randomized Control Group Pre & Post	18	Significant	None	Increase in emotional stability; however, change was of small practical significance
Higgins et al. 1981)	Education (preservice teachers)	B	B	Skill	Egan's scale	Randomized Control Group Post only	13	Significant	None	Increase in communication skills, and attitudes toward themselves as teachers and the ideal teacher
Layton (1979)	Nursing (nursing students)	B	B	Video Stimulus	Carkhuff Scale & Barrett-Lennard Relationship Inventory	Randomized Control Group Post &Follow-up	11	Significant	Age; Condition; Treatment by time	Increase in communication skills for junior students but not for senior students. Among the junior groups, only the groups receiving the rehearsal conditioning accounted for the significant group differences
Long et al. (1999)	Couples	B, C	C	Skill	Self Dyadic Perspective	Randomized Control Group	24 Couples	Significant	Time by gender	Increase in expression of empathy, ability to

Study	Discipline	Training Content*	Measure-ment Content*	Training Method	Measurement Instrument	Experimental Design	Sample Size	Significance of Findings	Moderating Variable(s)	Key Findings
					Taking Scale, Other Dyadic Perspective Taking Scale, IRI – Perspective-taking subscale	Pre, Post &Follow-up				understand the partner's point of view and greater satisfaction with the relationship
Long et al. (2006)	Couples	B	Unclear	Video Stimulus	Looked for precursors to empathy in participants	Qualitative Study	10 Couples	Irrelevant	None	After watching themselves on videotape couples reported to be surprised by self, experience discomfort watching the tape, increase in self-awareness and self-disclosure, and being able to observe patterns of interaction
McConnell & LeCapitaine (1988)	Education (secondary teachers)	Unclear	Unclear	Didactic & Experiential	California Psychological Inventory (Empathy); Teacher Behavior Description Questionnaire	Randomized Control Group Pre & Post	13	Significant	Age; Degree; Combination of experience, age, degree	Increased openness to student ideas and responses, allowing experimentation, and listening more intently
Nerdrum (1997)	Social Work (students)	B	B	Didactic & Skill	Carkhuff Scale	Control Group Pre, Post &Follow-up	29	Significant	Training and field practice	Increase in communication skills
Nerdrum & Ronnestad (2002)	Therapy (lay &professional therapists)	B, C	Unclear	Didactic & Experiential	Interview	Qualitative Study	23	Irrelevant	None	Credible improvement in empathic behavior in providing therapy but difficulty in internalizing and consolidating the learning.
Ridley et al. (1982)	Couples	B	A, C	Skill	Relationship Scale - Self &Relationship	Randomized Comparison	27 Couples	Significant	None	Increase in successful self-disclosure and role-taking

Study	Discipline	Training Content*	Measure-ment Content*	Training Method	Measurement Instrument	Experimental Design	Sample Size	Significance of Findings	Moderating Variable(s)	Key Findings
					Scale – Partner (based on Barrett-Lennard Relationship Inventory)	Group Pre & Post				abilities
Seto et al. (2006)	Therapy (counseling students)	C	A, C	Didactic & Experiential	IRI	Comparison Group Pre & Post	16	Non-significant	None	No change in the ability to relate to others' experiences and intolerance for ambiguity but a promising trend in the capacity to work in cross-cultural settings
Shapiro et al. (2006)	Medicine (medical students)	A, C	A, B, C	Writing	Ratings of written responses & observations	Randomized Comparison Group Post only	46	Significant	None	Increased awareness of emotional and spiritual aspects of a clinical encounter through writing
Shapiro et al. (1998)	Medicine (premedical &medical students)	B	Unclear	Mindfulness	Empathy Construct Rating Scale	Randomized Control Group Pre & Post	36	Significant	None	Reduced overall psychological distress, state and trait anxiety, and greater spirituality, listening skills and develop more compassionate perspectives towards themselves and others
Warner (1984)	Education (elementary & high school teachers)	B	B	Video Stimulus	Written responses to video	Randomized Control Group Post only	14	Significant	None	Increase in responding skills to anxious, depressed and angry students

extent to which the simulations or games induce motivation and/or the extent to which motivation contributes to empathy development (Ruben & Budd, 1975).

Didactic and Experiential Training. In *didactic and experiential* training, the facilitator lectures on theory and concepts and then provides experiences for the participants through games, internships, live cases, problem solving, and so on. Twenty-one percent of the evaluations we reviewed use the didactic and experiential training approach (Seto et al. 2006; Gantt, Billingsley, & Giordano, 1980; Avery & Thiessen, 1982; Nerdrum & Ronnestad, 2002; Aspy et al. 1984; McConnell & LeCapitaine, 1988).

With regard to research on the effectiveness of *didactic and experiential methods* to train people in empathy, Aspy et al. (1984) observed an improvement in teachers' interpersonal skills, including levels of empathy, and in classroom performance of the students they taught. Similarly, working with teachers, McConnell and LeCapitaine (1988) were able to increase teachers' levels of empathy, interactions with students, and openness to students' ideas and responses. Gantt et al.'s (1980) research findings showed that empathic sensitivity of mental health/human services students increased after a 10-week interviewing course. The authors noted that the improved empathic sensitivity was not only maintained but continued to increase over time. Avery and Thiessen (1982), who provided training in empathy and self-disclosure skills for divorcees, found that following training, the experimental group significantly increased their perceived level of social support, as well as self-disclosure and empathy skills. The authors noted a gender difference, with

females demonstrating a greater increase in empathy than their male counterparts.

Not all findings pertaining to effectiveness of *didactic and experiential* training are positive. Nerdrum and Ronnestad (2002) conducted a qualitative study to examine the trainees' conception of the learning process following participation in an empathy training program. The trainees reported considerable difficulty in trying to change their therapeutic style and argued that empathy was difficult to achieve. Using a quasi-experimental treatment-comparison groups design, Seto et al. (2006) investigated the effectiveness of a Triad Training Model in which the counselors are trained through an experiential exercise to identify and verbalize what the client is thinking, and they found no significant difference between groups.

Skill Training. *Skill* training is used in 17% of the studies we reviewed (Haynes & Avery, 1979; Herbek & Yammarino, 1990; Higgins, Moracco, & Danford, 1981; Long et al., 1999; Ridley, Jorgensen, Morgan, & Avery, 1982). This form of empathy training consists of three components: (1) provide trainees with a description of well-defined skills to be learned, (2) demonstrate the effective use of these skills through modeling, and (3) provide practice opportunities using these skills (Salas & Cannon-Bowers, 2001). Skill training may or may not involve trainers giving feedback to the trainees regarding their use of the skill.

In regard to *skill* empathy training research, Herbek and Yammarino (1990) studied the effectiveness of a *skill* training program for nurses and observed an increase, albeit of small practical significance, in the mean empathy scores for both the experimental and the control

group after training. Higgins et al. (1981) investigated empathic response skills of preservice teachers following either a group approach to human relations training (HRT) or a traditional approach to HRT. Findings suggested that both approaches increased preservice teachers' empathic scores and both approaches resulted in higher scores on teachers' attitudes toward themselves as teachers and toward the ideal teacher as well. In Haynes and Avery's (1979) study, the researchers found that students who participated in a communication skills training program demonstrated significantly higher self-disclosure and empathy skills levels than the control group. Based on this finding, the authors argue for teaching these skills to young populations.

Skill empathy training has also been used for couples' therapy. Long et al., (1999) observed that couples' empathy scores increased in both the treatment and wait listed comparison groups. While the authors did not find a significant gender difference, they noted that females showed a more rapid response to training over time. They also observed a positive relationship between change in empathic expression with a partner and relationship satisfaction six months after training. Ridley et al. (1982) assessed the effects of a relationship enhancement program in couples. The authors indicated that due to the training program, the couples reported significantly increased relationship adjustment; empathy, warmth and genuineness; trust; and couple communication.

Didactic and Skill Training. The combination of *didactic* and *skill* training (herein referred to as *didactic and skill* training) is used in 21% of the studies we reviewed (Ancel 2006; Crabb Moracco, &

Bender, 1983; Evans, Stanley, & Burrows, 1993; Fernandez-Olano et al. 2008; Hatcher, et al. 1994; Nerdrum 1997). With regard to research on the effectiveness of *didactic and skill* training in enhancing empathy, Fernandez-Olano et al. (2008) found that training slightly improved the empathy levels for medical students and medical residents. Women's empathy scores were higher than men's on both pretest and posttest. Likewise, Nerdrum (1997) reported that social work students' levels of communicated empathy were increased and maintained eighteen months after training. Hatcher et al. (1994) found that empathy can be successfully taught to college students as measured by how well they can empathize with fictional characters. However, high school students showed only a non-significant although positive trend, thus leading the authors to conclude that teaching students to become empathetic is more effective during the college than high school years. Working with medical students, Evans et al. (1993) observed no change in the participants' level of empathy using a pencil-and-paper test of empathy; however, independent observer ratings of participant behavior suggested an increase in empathy. Ancel (2006) found a significant increase in nurses' empathic skills but the increase was moderated by age and education levels. Empathy scores of nurses aged 31 years or older were lower than those of nurses 20-25 years old, and the scores of those with less than a Bachelor of Science education were lower than those with a Bachelor of Science. The findings of Crabb et al. (1983) indicated that both skill training alone and skill training combined with didactic training can significantly enhance the levels of empathy of lay persons in evangelical churches.

Mindfulness Training. Mindfulness training involves teaching trainees to become mindful, i.e., to be in a state of non-judgmental awareness grounded in the present moment. Being mindful is presumed to facilitate empathic responding skills. Ten percent of the studies we reviewed belong to this category of empathy training method. Research has shown that training in mindfulness helps people take other people's perspectives and feel more concern for them (Block-Lerner, Adair, Plumb, Rhatigan, & Orsillo, 2007). In mindfulness training, participants are directed to relax, to follow meditative practices and to think positively. Meditative practices aim to enhance participants' awareness of body sensations, sounds, thoughts, and emotions while continually focusing on the breath.

With regard to research on *mindfulness training*, preliminary results from Block-Lerner et al.'s study (2007) suggest that participants in the "mindful awareness" condition demonstrate a slight increase in their capacity to take other people's perspectives. Shapiro, Schwartz, and Bonner (1998) examined the effects of an 8-week, mindfulness-based, stress reduction intervention on stress and empathy in nursing students. Although participation in the intervention significantly reduced anxiety, it did not significantly increase empathy scores. Similarly, Beddoe and Murphy (2004) found a reduction in participants' anxiety but did not observe an increase in empathy.

Apparently the effect of mindfulness on empathy may be an indirect one. Being mindful helps individuals become more at ease and less critical, which can lead to a sense of compassion and consequently a feeling of empathy towards others. One can argue that mindfulness training is really not an empathy training method since it does not directly address an individual's empathy. The relationship between mindfulness and empathy warrants further consideration, and the use of mindfulness training to induce empathy requires further conceptualization and research**.**

Video Stimulus Training. In *video stimulus* training the trainer asks the participants to watch a videotape about others' empathic behaviors, or their own, in mock situations, and to respond to the videotaped excerpts during the viewing or afterwards. The training session could also be followed by discussion and feedback. Seventeen percent of the studies we reviewed employed videotape training (Barone et al., 2005; Fraser & Vitro, 1975; Layton, 1979; Long et al., 2006; Warner, 1984).

In one of these studies (Barone et al., 2005), the participants were asked to infer thoughts and feelings of the subject in the tape. Meanwhile, Long et al. (2006) implemented a qualitative study to investigate the effects of videotaping couples' interactions in a relationship enhancement program, and to explore how the viewing of the videotape could be a useful technique for enhancing empathy. Unfortunately, they did not draw any clear conclusion about the process and effects of the intervention because their study was exploratory and they were unable to account for confounding factors in their research. Warner (1984) found an increase in teachers' empathic responses towards students as a result of video training. Layton (1979) who used various combinations of modeling, labeling, and rehearsal (videotaped) to teach empathy to nursing students, found that the

treatment was effective for junior but not for senior students. In addition, only the groups receiving the rehearsal condition (10-second pauses following each videotaped client statement where participants were asked to mentally construct their own responses) performed better than the control group. Barone et al. (2005) investigated the effects of *video stimulus* training on psychology students' empathy. Their findings showed that participants in the experimental groups were more accurate than the control subjects, and women were more accurate than men in inferring thoughts. Fraser and Vitro (1975) reported highly significant increases in empathic response in their experimental group. However, the authors revealed that neither the experimental nor control group reached the minimally facilitative empathy level as described by Carkhuff (1969).

Writing Training. Writing training, as the name implies, is a training method that entails asking trainees to write from the other's point of view or perspective as the strategy for enhancing empathy. This particular method was used in two (7%) of the 29 research evaluations we reviewed and both studies (DasGupta & Charon, 2004; Shapiro, Rucker, Boker, & Lie, 2006) evaluated training for medical students. In one of the two studies, participants were asked to write about their personal illness or a relative's illness, while in the other study they were asked to write from a patient's perspective, referred to as point-of-view writing. The qualitative analysis of DasGupta and Charon (2004) indicated that the training was well-received by medical students even though they struggled with writing about their personal illness experiences. Although medical educators use writing as a method to enhance medical students' empathy, there is a lack of evidence supporting transfer of learning to clinical practice (Shapiro et al. 2006)**.**

Summary. Figure 1 summarizes the percent of studies that examined each of these seven training methods. As seen in Figure 1, the most commonly used training method is the combination of didactic and experiential training and the combination of didactic and skill training; the least used are experiential training and writing.

With regard to findings about training effectiveness, all but two of the 26 (92%) evaluations that used quantitative methods reported positive findings regarding the trainability of empathy. Among the three qualitative evaluations, two studies (Long et al., 2006; Nerdrum & Ronnestad, 2002) did not provide conclusive statements about the effectiveness of training. Since we considered findings from these studies as statistically non-significant, only one of the three (33%) qualitative research studies reported significant findings. Collectively, findings from the 29 studies lead us to conclude that empathy is generally trainable. (See Table 1 for a summary of key findings from each study.)
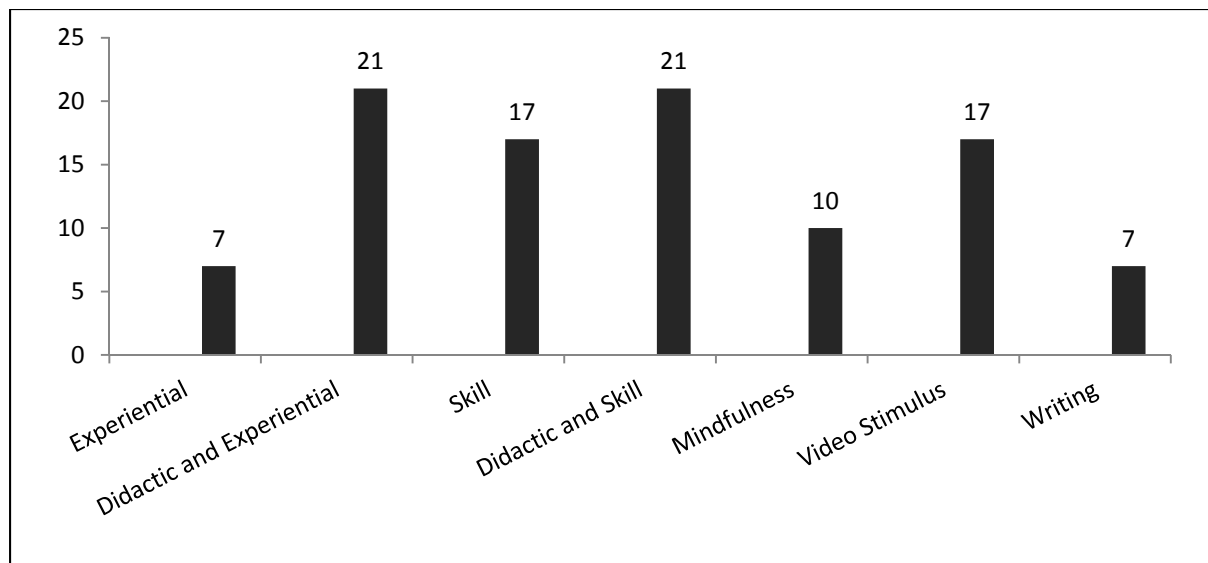
Figure 1. Percent of Training Methods Used in the 29 Studies Reviewed

The next logical question we wanted to address was, "Which of the three types of empathy is trainable?" To do that, we delved deeper into our analysis and examined the definition of empathy, and the training and measurement content of the 29 studies we reviewed.

We were surprised that 10 of the 29 studies we reviewed did not provide a definition of the nature of empathy investigated in their research (Aspy et al., 1984; Avery & Thiessen, 1982; Crabb et al., 1983; Fraser & Vitro, 1975; Gantt et al., 1980; Higgins et al., 1981; McConnell & LeCapitaine, 1988; Ridley et al., 1982; Seto et al.,2006; Shapiro et al. 2006). Consequently we resorted to examining the training and measurement content to infer the type(s) of empathy each of these studies was targeting to affect. In Figure 2, we depict the percent of studies that focus on the different individual and combination of empathy components as determined by the training content and the measurement content. (The training and measurement content for each individual study can be seen in Table 1).
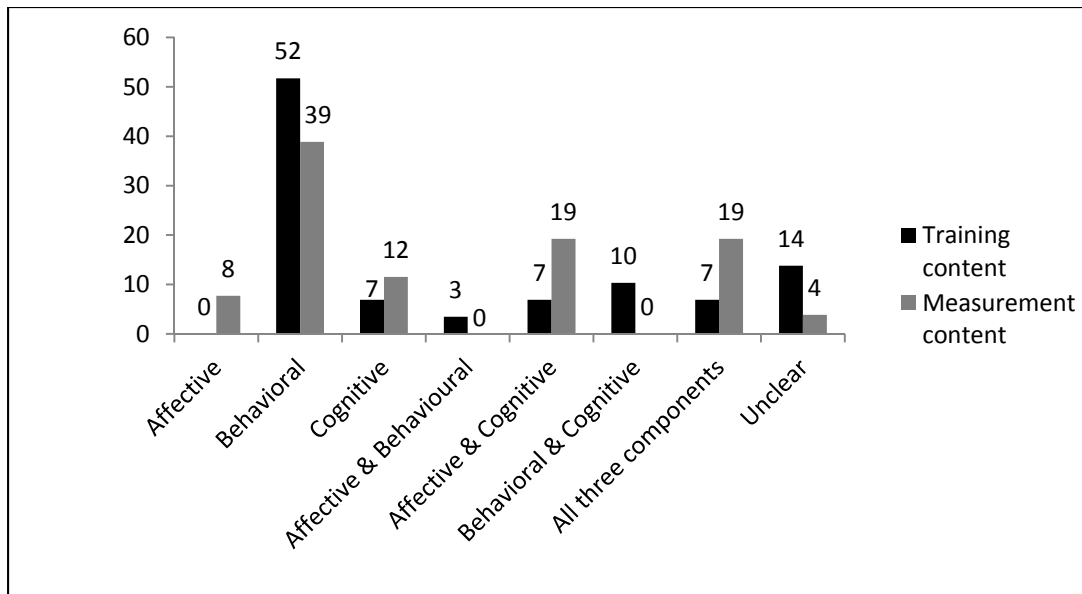
Figure 2. Percent of Qualitative and Quantitative Studies with Training Content of Affective, Behavioral, or Cognitive Aspects of Empathy, and Percentage of Quantitative Studies with Measurement Content of Affective, Behavioral or Cognitive Aspects of Empathy

On the basis of training content, we found that most of the studies (52%) focused solely on training participants to behave empathetically. Our analysis of the measurement content also indicated that the majority of the studies were targeting the behavioral aspect of empathy (39%). With this observation, we elaborate on the empathy trainability finding and conclude that irrespective of the training method, empathy is trainable, especially the behavioral component of empathy. It appears that we can train people to act and communicate empathetically, although whether empathetic behavior is accompanied by the ability to understand others' perspectives and the feeling that is associated with such understanding is not clear.

When the target empathy components were inferred on the basis of the training content, the distribution of the 29 studies across the individual and combined empathy components that these studies examined was not the same as when they were inferred from the measurement content (see Figure 2). Some studies we reviewed did not measure what they trained. This discovery of a seeming mismatch between training and measurement of empathy, together with a lack of empathy construct explication, leads into the next part of our paper that addresses the question: How was the empathy training evaluated and how valid are the findings from these evaluations?

Evaluation Practices and Validity of Findings

Gurman and Kniskern (1978) reviewed over 200 studies on the effects of marital and family therapy. Besides counting the outcomes, they rated the design quality of each study based on the criteria such as controlled assignment to treatment

conditions, pre-post measurement of change, appropriate statistical analysis, multiple change indices used, and so on. We also conducted a similar analysis of the quality of research designs or practices employed in the 29 studies we reviewed based on the information made available to us in the articles and the fourfold validity framework proposed by Shadish, Cook & Campbell (2002). We examined some aspects of or threats to (1) construct validity of the training, the measurements of empathy and trainee samples; (2) statistical power or conclusion validity; (3) internal validity of the findings regarding impact of training on the participants; (4) generalizability of findings to other constructs or external validity.

According to Shadish, Cook & Campbell (2002), construct validity refers to "the validity of inferences about the higher order constructs that represent sampling particulars [pertaining to persons, settings, treatments, and outcomes]" (p.38). Consequently, "threats to construct validity concern the match between study operations and the constructs used to describe those operations" (p.72). Generally, construct validity is low if either the operationalization fails to incorporate all the prototypic characteristics of the target construct (construct underrepresentation) or contain content extraneous to the construct (irrelevant construct content).

Our examination of construct validity of the training implemented by the 29 studies entailed examining the extent to which the researchers delivered the training that was consistent with how they defined it. We determined construct validity of the outcomes by examining (1) the extent to which the instruments or procedures used in the 29 studies to measure empathy indeed measured empathy (for which we reviewed the psychometric information provided by the studies); (2) the extent to which the researchers measured what they trained (for which we examined the match between measurement content and treatment content, since one-third of the studies we reviewed did not define empathy); (3) the extent to which the researchers collected data about how much of the training had been transferred to the natural environment in which the participants were expected to behave empathetically on their own volition. We determined construct validity of persons by examining the extent to which the researchers examined trainee samples' representativeness of the target population.

Beyond the concerns regarding validity of the training, empathy and participant constructs is whether the 26 quantitative studies had sufficient power in their statistical analyses to detect treatment effects (i.e., statistical conclusion validity), and the training did in fact cause the observed change in post-training empathy of the participants (i.e., internal validity). We determined validity of conclusions from statistical analysis by examining measurement reliability and sample size, and internal validity of the impact findings by reviewing the quantitative studies' experimental designs and the flaws embedded in the experimental procedures. Finally, with regard to external validity or the extent to which researchers drew correct conclusions about generalizability of findings to other constructs, we checked if the 29 studies examined effects of moderating variables.

We developed a rating scheme system consisting of 14 criteria. Five of these criteria were not appropriate for qualitative research and consequently we did not include the three qualitative

studies in our rating of these criteria; these five criteria are measurement validity and reliability, sample size, use of control group, and use of pre-training measures. The 14 criteria and their associated rating scales are shown in Figure 3.

Figure 3
Criteria and Scales for Rating Quality of Methodological Practices.

Criteria	Rating Scales		
	1	1	0
	1	0.5	0
A. Construct validity of treatment			
A1. Training method definition stated		Yes	No
A2. Description of training provided		Yes	No
A3. Congruence between training method definition and training implementation description	Complete	Partial	None/Un clear
B. Construct validity of empathy measurement			
B1. Validity	Measured	Reported*	No
B2. Definition of empathy provided		Yes	No
B3. Congruence between measurement content and training content	Complete	Partial	None/Un clear
B4. Congruence between empathy definition, training content and measurement content	Complete	Partial	None/Un clear
B5. Highest level of outcomes measured (Kirkpatrick's outcome levels)	Transfer or Results	Learning	Reaction
C. Construct validity of persons			
C1. Examination of representativeness of trainee sample	Yes	Partial	No
D. Statistical conclusion validity**			
D1. Reliability**	Measured	Reported*	No
D2. Sample size**		= or > 25	<25***
E . Internal validity (Research design) **			
E1. Use of control group**	Yes, check group comparability or random assignment	Yes	No
E2. Use of pre-training measure****		Yes****	No
F. External validity			
F1. Examination of moderator variable effect		Yes	No

*Reported from literature.
**Criteria do not apply to qualitative studies.
***Rating of 0 is also assigned to studies that did not report sample size.
****Pre-training testing or retrospective pre-testing.

It should be noted that for a criterion with a two-point rating scale (yes or no), the average % criterion attainment score simply refers to the percent of studies that had met the criterion. For a criterion with a three-point rating scale (yes, partial, no), the average % criterion attainment score refers to the degree to which the criterion was reached by the studies included in the analysis, and not the percent of studies that had met the criterion. In both cases the average % criterion attainment score ranges from zero to 100.

To address the second question of our review, "How was empathy training evaluated and how valid are these evaluation findings?" we referred to our ratings of the aforementioned design criteria and used the following specific questions to guide our responses:

- To what extent have researchers implemented training in accordance to their definition of the training method?
- How have researchers measured empathy and how psychometrically sound are these measurements?
- To what extent have researchers measured what they trained?
- To what extent have researchers measured transfer of learning?
- How have researchers measured the impact of training on empathy and how valid and generalizable are these estimates?

The first four questions relate primarily to construct validity, and the last question, statistical conclusion validity, internal validity, and external validity. Next we will present our findings in regard to these five questions.

To What Extent Have Researchers Implemented Training in Accordance to Their Definition of the Training Method?

In order to assure the construct validity of empathy training, it is necessary for the researchers to explicate their treatment constructs, and to gather evidence in support of the treatment fidelity (Sechrest et al., 1979; Summerfelt, 2003). To determine the extent to which researchers of the 29 studies we reviewed provided training according to how they conceptualized their training method, we searched for the researchers' definition of the training method they used, their description of how the training was delivered, and the process data they collected to support their training delivery description. In regard to gathering actual implementation data (e.g., the actual amount of time the trainer devoted to different aspects of the training or used different strategies), we found only one study, Higgins et al. (1981), that collected such data. For example, these authors reported that "about 50% of the time was spent on empathy training [and] none of the first seven sessions in the A approach was spent on developing specific skills in human relations training" (p. 23). Since only one study reported implementation data, we judged the congruence between treatment conceptualization and implementation on the basis of the match between training procedure implied by its definition (augmented by our knowledge of the training method) and the researchers' description of how they delivered training. Our analysis led us to the following three types of studies.

Studies with Congruence between Definition and Implementation of Training Methods

Eighteen of the 29 studies (62%) that we reviewed had congruence between the researchers' definition of their training methods and their description of how they delivered their training, including the only study that collected and reported implementation data by Higgins et al. (1981). These studies are: Ancel, 2006; Barak et al., 1987; Barone et al., 2005; Beddoe and Murphy, 2004; Block-Lerner et al., 2007; Crabb et al., 1983; Fernández-Olano et al., 2008; DasGupta and Charon, 2004; Gantt et al., 1980; Higgins et al., 1981; Layton 1979; Long et al., 1999; Long et al., 2006; Nerdrum, 1997; Seto et al., 2006; Shapiro et al., 2006; Shapiro et al., 1998; and Warner, 1984.

It should be noted that researchers in three studies (10%) (Beddoe & Murphy, 2004; Block-Lerner et al., 2007; and Shapiro et al., 1998) based the theoretical rationale for training effectiveness on mindfulness instead of empathy. As indicated earlier, the relationship between empathy and mindfulness is unclear, so the theoretical rationale for how the training is expected to increase empathy is questionable. Beddoe and Murphy (2004) conceded that re-designing a mindfulness-training program "by adding an 'empathy towards others' component through specific meditations, poetry reading, and group discussion" (p. 309-310) should be more effective in enhancing empathy. However, all three studies appeared to be consistent in defining and in describing how the training was delivered, and hence we considered all these studies as having congruence between treatment definition and treatment implementation description.

Studies with Partial Congruence between Definition and Implementation of Training Methods

One study (3%) had partial congruence due to the omission of training element(s) in the training implementation description. Nerdrum and Ronnestad (2002) defined their training as both didactic and experiential, but they did not refer to the experiential component when describing their training.

Studies with No Congruence between Definition and Implementation of Training Methods

Ten studies (34%) had either no congruence between training definition and implementation description or did not include enough information for us to determine the congruence and hence construct validity of the training for these studies (two quantitative studies did not define the training method and seven quantitative studies did not provide enough detail on training delivery). These ten studies are: Aspy et al., 1984; Avery and Thiessen, 1982; Evans et al., 1993; Feighny et al., 1995; Fraser and Vitro, 1975; Hatcher et al., 1994; Haynes and Avery, 1979; Herbek and Yammarino, 1990; McConnell and LeCapitaine, 1988; Ridley et al., 1982.

Overall Findings

As mentioned above, for evaluating training implementation, not only must

researchers explicate their treatment constructs, they must also collect data to show how the training was actually delivered. Such process data are important to assure treatment fidelity or high treatment construct validity as discussed. Although we found 18 (62%) the 29 studies we reviewed had congruence between treatment definitions and treatment implementation descriptions, only one of these 18 studies provided process data in support of their descriptions. Apparently, data based treatment implementation descriptions have greater credibility than non-data-based treatment implementation descriptions. In fact, describing treatment implementation without data is no different than operationalizing the treatment definition. Although an important process, it is not sufficient to validate a treatment construct because there is no evidence demonstrating the extent to which operationalization has been implemented. Of the remaining 11 studies, one had a partial match while in the rest, the treatment definitions and implementation descriptions did not match. These findings are depicted in Figure 4.
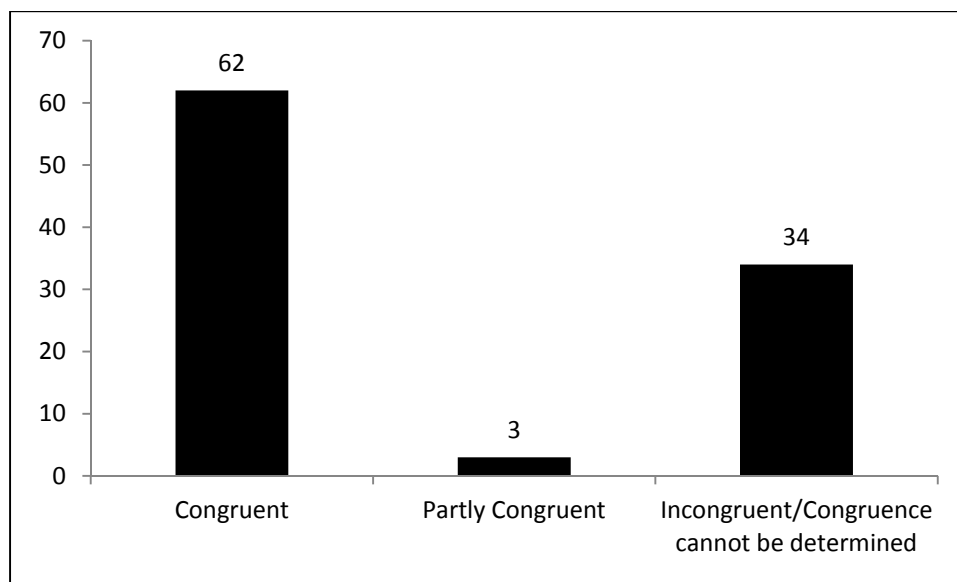

Figure 4. Percent of Studies that have Congruence, Partial Congruence, or No Congruence between Training Definition and Implementation Description.

With regard to rating the studies' design, we noted 90% of the studies stated a training method definition, 76% provided a description of the training implementation, and we assigned the 29 studies an average 64% criterion attainment score for the congruence between training method definition and training implementation description.

In sum, we feel that the treatment construct validity of findings generated by the studies we reviewed was collectively rather low. Some researchers in these studies failed to clearly define their training methods, some failed to describe

how the training was delivered, some failed to provide training as planned, and almost all failed to present evidence to show how they delivered their training.

How have Researchers Measured Empathy and How Psychometrically Sound Are These Measurements?

The majority of the studies included in this analysis (25 out of 29 studies including the three qualitative studies, or 86%) used trainees' written responses to measure empathy, which could be responses to either open-response questions or closed-response questions. Eight of these studies used open-ended writing tasks (six of which used open-response questions exclusively), while 17 studies used closed-response questions (14 of which used closed-response questions exclusively). Open-response questions were usually vignettes that describe a situation through either a written description of an event or a videotaping of a scenario, and trainees are asked to describe how they would respond to the person in that situation. Closed-response questions were usually in the form of a Likert-type rating scale. We noted a total of eight scales were used in the studies we reviewed. The most common empathy scale used was Davis' Interpersonal Reactivity Index (IRI) scale; seven studies used this particular scale.

Another method of measuring empathy used in some of the studies was observation and rating of trainees' interactions (use of words and tone) with another person either in a real-life situation or via an audiotape recording of the interactions. Six studies used this measurement method, three of which used closed-response scales as well, and two used open-response questions, in addition to the observer ratings. Only one study (Barone et al., 2005) used observer ratings as the only form of measurement of empathy. The frequency of use of the various methods to measure empathy by the 29 studies is depicted in Figure 5. (In Table 1, we show the measurement instrument[s] used in each of the 29 studies we reviewed).

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C. Alamparambil
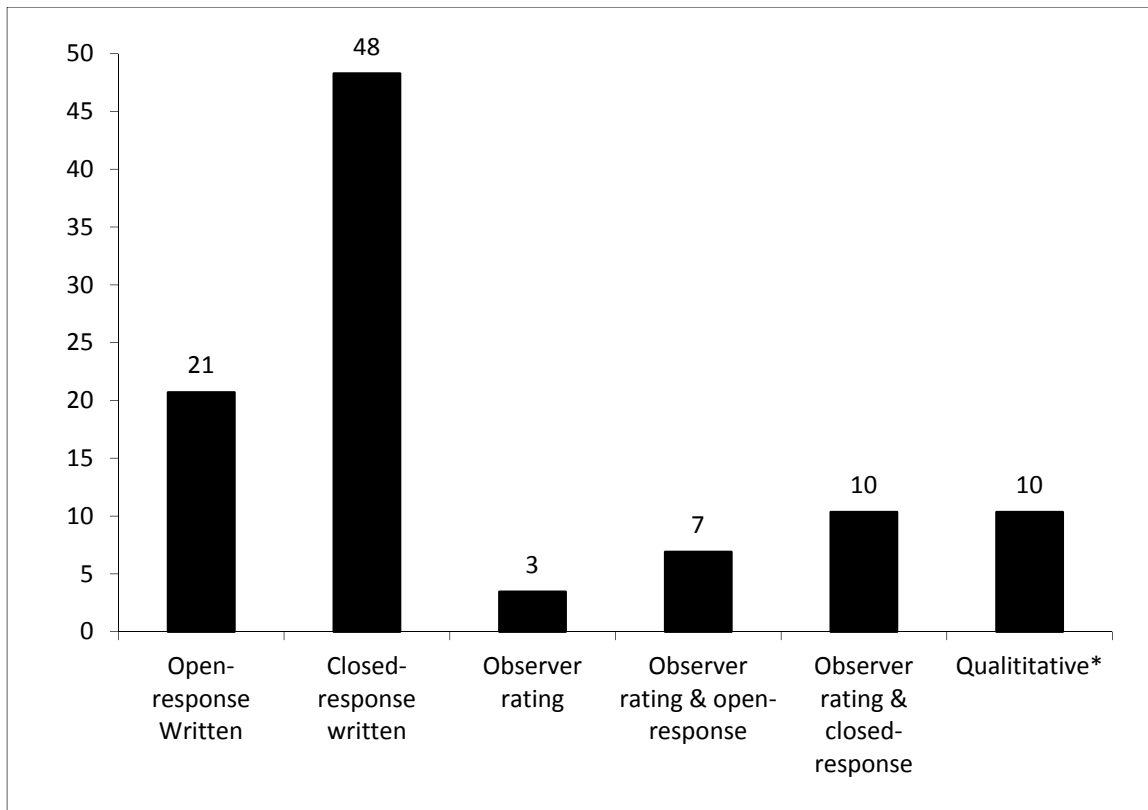

Figure 5. Percent of Quantitative Studies Using Open-Response Written Tests of Empathy, Closed-Response Written Tests, Observer Ratings, or More than One of the Above

*The "qualitative" category comprises the studies that used qualitative methods.

Reliability and Validity of Empathy Measurements. Reliability. Nine (34%) of the 26 quantitative studies (Ancel, 2006; Aspy et al., 1984; Barak et al., 1987; Block-Lerner et al., 2007; Crabb et al., 1983; Feighny et al., 1995; Fernandez-Olano et al., 2008; McConnell & LeCapitaine, 1988; Nerdrum, 1997) did not provide data about reliability of the measurement instruments used in the studies. One study (Ridley et al., 1982) referred readers to other sources to look up reliability information.

Of the remaining 16 studies, 7 studies (Avery & Thiessen, 1982; Barone et al., 2005; Fraser & Vitro, 1975; Haynes &

Avery, 1979; Higgins et al., 1981; Shapiro et al., 2006; Warner, 1984) reported inter-rater reliability of observer ratings. Five studies reported internal consistency estimates (as either Cronbach's alpha, split-half reliability), and one study provided test-retest reliability. Two studies (Evans et al., 1993; Layton, 1979) used multiple measurement instruments, and reported estimates on inter-rater reliability, internal consistency, and test-retest reliability. One study (Hatcher et al., 1994) reported both the internal consistency and test-retest reliability of their instrument.

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C. Alamparambil

The inter-rater reliability reported in these studies ranged from 0.36 to 0.95, the internal consistency scores ranged between 0.68 and 0.88 and the test-retest reliability was between 0.62 and 0.89. All of these reliability estimates were reasonably high (0.62 or more) except for those described by Evans et al. (1993) who reported inter-rater reliability of 0.36 to 0.39 for the Accurate Empathy Scale. The authors explained that the observed level of reliability was to be expected because their reliability coefficients were rank-order correlation statistics (Spearman's rho). According to our rating scheme, the percent criterion attainment score for measurement reliability is 63%.

Validity. Only eight (31%) of the 26 quantitative studies address the issue of validity of their measurements (Block-Lerner et al., 2007; Evans et al., 1993; Hatcher et al., 1994; Herbek & Yammarino, 1990; Layton, 1979; Long et al., 1999; McConnell & LeCapitaine, 1988). Of the eight studies, only two studies reported evidence of construct validity of the empathy measurements developed for their own studies. The other six provided validity evidence reported from previous studies that used the same empathy measurement instrument.

Block-Lerner et al. (2007) examined both convergent and discriminant construct validity of their empathy scores by correlating them with those of two mindfulness scales, Cognitive and Affective Mindfulness Scale-Revised (CAMS-R) and the Mindfulness Attention and Awareness Scale (MAAS). There was a significant correlation between scores on two of the subscales of the Interpersonal Reactivity Index with scores on the CAMS-R (r = 0.33 to 0.35). However, there was no significant correlation between the IRI and MAAS scores.

Layton's study likewise gathered convergent and discriminant construct validity evidence by comparing scores from their written test that measures knowledge about empathy (Layton's Empathy Test) with 3 empathy instruments: Carkhuff's rating scale (Carkhuff, 1969), the Barrett-Lennard Relationship Inventory (Barrett-Lennard, 1962) and LaMonica's Empathy Construct Rating Scale (LaMonica, 1976). Layton's written empathy test scores significantly correlated with ratings from the Carkhuff scale (r = 0.25), but not with either one of the other two instruments.

Researchers of the remaining six studies simply referred to or presented findings from previous validation studies conducted on the instruments they used in their studies. Herbek and Yammarino (1990) quoted findings from Mehrabian's studies of intercorrelation of subscales and correlation to helping behavior (Mehrabian & Epstein, 1972). Hatcher et al. (1994) quoted Davis' validation studies (Davis & Franzoi, 1991), which showed convergent as well as discriminant validity of the subscales of the IRI. McConnell and LeCapitaine (1988) referred to Baucum's (1985) research that showed construct validity of the CPI (California Psychological Inventory) scales. Long et al. (1999) stated that Davis and his collaborators conducted validity studies on the IRI, but did not describe the studies. For their other empathy measures, the Self Dyadic Perspective Taking scale (SDPT) and Other Dyadic Perspective Taking scale (ODPT), Long et al. (1999) simply stated, without further elaboration, that "an item analysis and assessments of the construct and concurrent validity indicated that the scale had adequate validity" (Long et al., 1999, p. 238). Evans et al. (1993) indicated the IRI's construct validity

evidence could be surmised from the fact that females generally score higher than males on the IRI as they do on other empathy measurement instruments. Ridley et al. (1982), the last of the six studies that referred readers to other studies for validity information, did not describe the type of validity information available in those sources.

In sum, most of the 29 studies did not gather evidence to examine the validity of their measurements, and the majority of the studies that reported any information on validity were reporting findings from other studies. The percent criterion score for measurement validity is 19%. These observations lead us to wonder about the extent to which the instruments used to measure empathy in the studies we reviewed are, in fact, measuring empathy as conceptualized by the authors for their studies. Thus, we delve further to examine the match between what the researchers measured and the form of empathy that they aim to enhance through the training.

To What Extent Have Researchers Measured What They Trained?

We noticed that only 66% of the 29 studies we reviewed provided a definition of empathy for their studies. However, even for studies that do not have a stated definition of empathy, if their measurement content and treatment content are congruent, they still have some degree of construct validity of measurement (or construct validity of treatment depending on one's perspective). We adhered to this premise when we conducted our review of the selected articles on construct validity of empathy measurement. We first analyzed the training and measurement content of all the studies, then we determined if a study had full, partial or no congruence between these two content types, that is, the extent to which a study measured the same aspects of empathy as they trained. Lastly, for the studies that had a stated definition of empathy, we determined if the definitions was consistent with the training and measurement content.

Congruence between Training and Measurement Content. We determined the training content in each of the 29 studies by reviewing the authors' description of the procedures and materials used in the training; similarly, we inferred the empathy component(s) measured by examining the content description of the instruments used to measure empathy. Then, we examined the match between the components of the empathy trained and those measured. Figure 6 depicts the frequency of studies that have full, partial, or no congruence between training and measurement content. (In Table 1, we present the training and measurement content for each of the 29 studies we reviewed.)
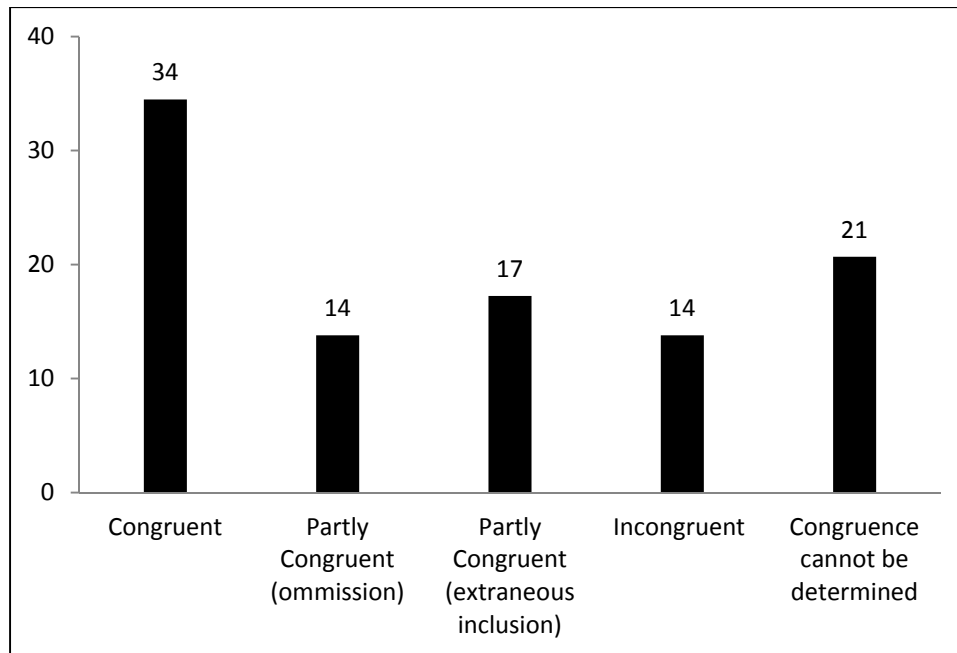
Figure 6. Percent of Quantitative Studies that have Congruence, Partial Congruence, or no Congruence between Training Content and Measurement Content

*Some studies have partial congruence by omission if they do not measure all aspects of empathy that they trained; or partial congruence by extraneous inclusion if they measure aspects of empathy that they did not train.

Full Congruence between Training Content and Measurement Content. Ten studies were consistent in the aspects of empathy that they trained and measured. Eight of them were consistent in that they trained participants in behavioral empathy alone, and measured solely behavioral empathy. Barone et al. (2005) trained participants on cognitive empathy and measured cognitive empathy. Feighny et al. (1995) trained and measured all three aspects of empathy.

Partial Congruence because of Omission (construct underrepresentation; Shadish et al., 2002, p.72). Four studies (Barak et al., 1987; Crabb et al., 1983; Herbek & Yammarino, 1990; Long et al., 1999) provided training to participants on more than one aspect of empathy, but measured only one of these aspects. These researchers, therefore, only measured the effectiveness of part of their training.

Partial Congruence because of Extraneous Inclusion (irrelevant construct content; Shadish et al, 2002, p. 72). Five studies measured aspects of empathy on which they did not provide training. Three of the studies (Ancel, 2006; Fernandez-Olano et al., 2008; Shapiro et al., 2006) measured all three components of empathy although the researchers only trained one or two. Layton (1979) trained participants on behavioral empathy, but measured both behavioral and cognitive empathy. Seto et al. (2006) trained participants on cognitive empathy, but measured both cognitive and affective empathy. Researchers in these studies did not provide a rationale for measuring the

component of empathy that they did not train.

Complete Incongruence. The four studies (Beddoe & Murphy, 2004; Block-Lerner et al., 2007; Hatcher et al., 1994; Ridley et al., 1982) that had no congruence between training content and measurement content invariably trained behavioral empathy but actually measured affective and cognitive empathy. It should be noted that the studies by Beddoe and Murphy (2004) and Block-Lerner et al. (2007) were problematic to categorize because rather than empathy training, they provided training on *mindfulness,* or a person's ability to maintain non-judgmental, present-minded awareness. We consider the empathy outcome of mindfulness training as behavioral because it involves a set of behaviors that the trainees learned to perform.

Congruence cannot be Determined. Six studies did not include enough information for us to determine either the training or measurement content. Evans et al. (1993) did not adequately describe the training content, and Shapiro et al. (1998) did not describe their measurement scale, the Empathy Construct Rating Scale, well enough for us to infer the type of empathy measured. McConnell and LeCapitaine (1988) used a problem-solving experience as training; consequently, it was hard to categorize the

component of empathy being trained. The three other studies were qualitative (DasGupta & Charon, 2004; Long et al., 2006; Nerdrum & Ronnestad, 2002) and did not measure any specific components of empathy.

In our rating of design quality in terms of the match between measurement content and training content in each of the 29 studies we reviewed, we arrived at a criterion attainment score of 62% based on the number of studies that had full, partial or no congruence. Next we determined the number of studies that had congruence not just between measurement content and training content, but also with the empathy definition as well.

Congruence between Definition of Empathy, Training Content, and Measurement Content. To further understand the quality of the research on the trainability of empathy, we attempted to isolate those studies that have a sound design, i.e., where the definition of empathy stated is consistent with the aspects of empathy that are trained and those that are measured. We analyzed the ten studies that had complete congruence between training and measurement content to determine if their stated definition was also consistent with the rest of the study. The results are summarized in Figure 7.
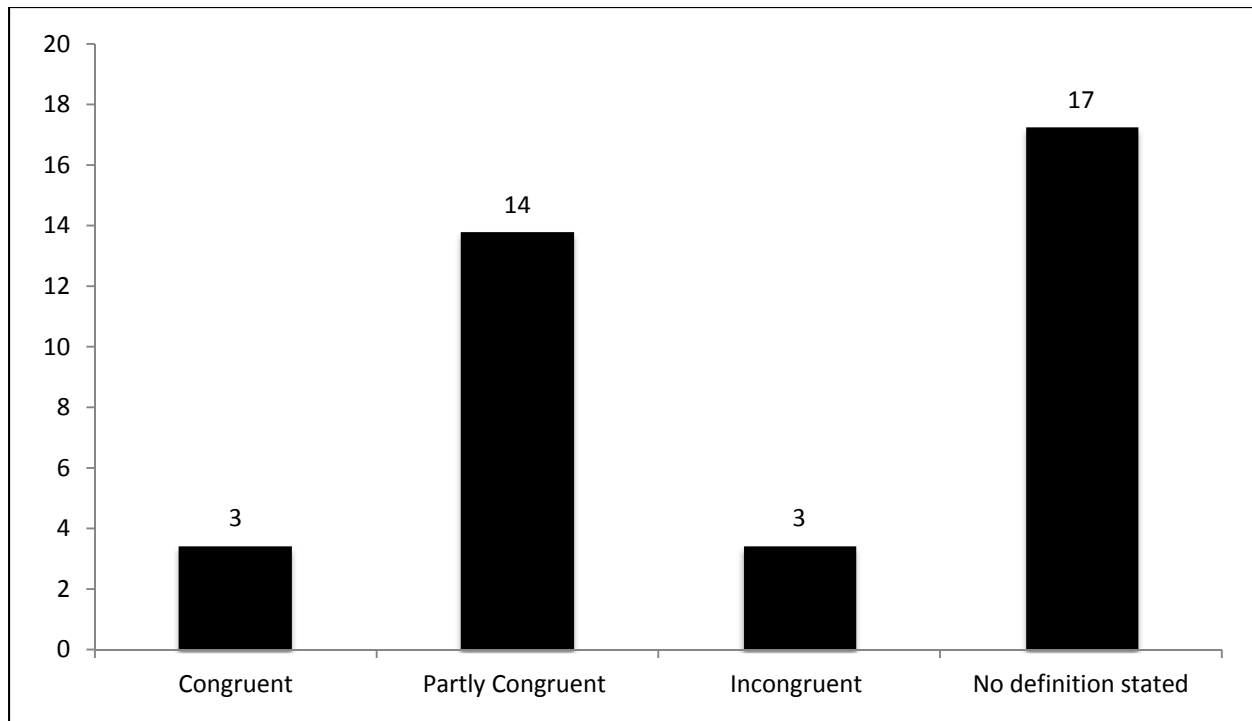
Figure 7. Percent of Quantitative Studies that have Congruence between Training and Measurement Content that were also found to have Congruence, Partial Congruence or no Congruence with Their Stated Definition

*Some studies had no definition stated although they had complete congruence between measurement and training content.

As can be seen in Figure 7, five studies (Aspy et al., 1984; Avery & Thiessen, 1982; Fraser & Vitro, 1975; Gantt et al., 1980; Higgins et al., 1981) that had congruence between training and measurement content did not provide a definition of empathy. Nerdrum (1997) trained and measured behavioral empathy, but defined empathy as a cognitive ability.

Researchers in three studies included more empathy components in their definitions than those they trained and measured. For example, Warner (1984) and Haynes and Avery (1979) were consistent in training and measuring behavioral empathy, but they defined empathy as being cognitive as well as

behavioral. Barone et al. (2005) trained and measured cognitive empathy, but defined empathy as having all three components.

We found only one study having full congruence between definition and training and measurement content. Feighny et al. (1995) were completely consistent in how they defined empathy and what they trained and measured. In their study, they defined empathy as having affective, behavioral and cognitive components, and they trained and measured all three components. From our rating of design quality, we assigned a low 9% criterion attainment score for congruence between empathy definition,

measurement content and training content.

We want to note that the study by Herbek and Yammarino (1990) was the only study in which the researchers explicitly indicated a conscious effort to be consistent in how they conceptualized and measured empathy. They acknowledged that empathy has cognitive and affective aspects, but they chose to focus solely on the affective one because they believed that it is the key component that enables nurses to be empathic to their patients. While they trained both affective and behavioral aspects of empathy, they measured only the affective component. No other study we reviewed provided reasons for omitting aspects of empathy from either their training content or measurement content.

We also want to acknowledge that construct validity threats are numerous (see Shaddish, Cook & Campbell, 2002) and given the limited data available to us, we were not able to identify the specific threats that were responsible for the observed mismatch between definition, measurement content, and training content. Some authors of the articles we reviewed suggested the possibility of such measurement construct validity threats as Hawthorne effect (McConnell & LeCapitaine, 1988) and researcher bias: some researchers who collected research data were either involved in the training

development (Dasgupta & Charon, 2004; Nerdrum & Ronnestad, 2002) or in the training itself (Crabb et al., 1983).

To What Extent Have Researchers measured Transfer of Learning?

Empathy training is only useful if, in fact, the trainees become more empathetic in their daily lives or in their work places depending on the purpose of the training. Hence, to understand the effectiveness of training, it is imperative that we examine the degree to which findings from the studies we reviewed reflect transfer of learning from training to situations in which empathy is required. This transference corresponds to behavior change or the Level-3 outcome in Kirkpatrick's (1994) classification of training outcomes. The other three levels are: Level 1: reaction (trainees' opinion about the training), Level 2: learning (how much trainees have learned), and level 4: results (the impact on the organization or clients). Evidence pertaining to transfer of learning has also been referred to as evidence for ecological validity (Bracht & Glass, 1968). Figure 8 shows the number of studies out of the 29 studies we reviewed that measured transfer of learning as well as outcomes at the other levels or combinations thereof.
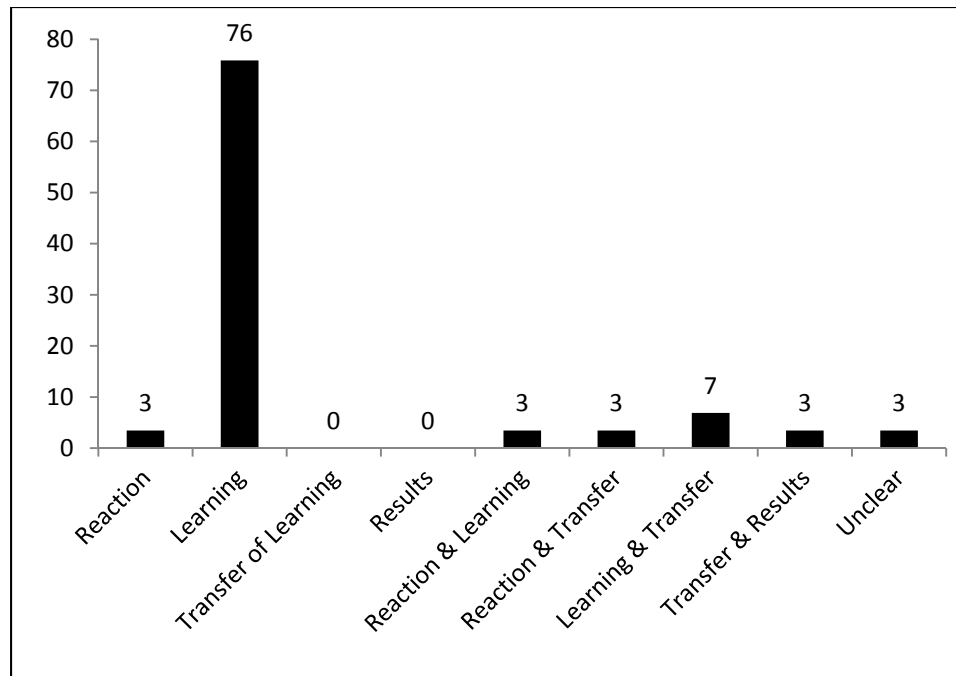
Figure 8. Percent of Quantitative & Qualitative Studies that Measured Different Individual and Combinations of Kirkpatrick's Four Levels of Training Outcomes

Of the three qualitative studies in our database, one (Long et al., 2006) did not measure any of the outcomes; one study (DasGupta & Charon, 2004) measured both reaction and transfer of learning using self-reporting; one study (Nerdrum & Ronnestad, 2002) documented participants' reaction to the training. As seen in Figure 8, the vast majority of the quantitative studies measured trainees' learning using a written test or a simulated interaction. Only one study (Ancel, 2006) measured both reaction and learning. Two studies (Beddoe & Murphy, 2004; Evans et al., 1993) measured both learning and transfer of learning, and one of these (Beddoe & Murphy, 2004) used self-reporting to measure transfer of learning.

There is only one study (Aspy et al., 1984) that measured expected training impact on clients together with transfer of learning. Aspy et al. (1984) provided empathy training to teachers, and then measured their transfer of empathy to the workplace by rating audiotaped interactions of teachers with their students in the classroom. These researchers also obtained findings showing that the students of empathy-trained teachers had better attendance rates, and higher achievement scores in certain subject areas.

Only four of the 29 studies (14%) we reviewed measured changes in empathy in the natural environment and two of these studies relied on self-reporting. Most of these 29 studies reported improvement of learning or acquiring empathy, a very small percentage measured changes in transfer of learning about empathy, and no studies examined the affective component of empathy. According to our rating scheme, we assigned a criterion attainment score of 59% to the 29 studies reviewed based on the level of outcomes

these studies measured (as shown in Figure 3, we assigned 1 point for measuring transfer or results, 0.5 point for learning, and 0 point for reaction).

How Have Researchers Measured the Impact of Training on Empathy and How Valid and Generalizable are These Estimates?

Thus far, the methodological findings we have presented pertain to the conceptualization and measurement of empathy itself and its congruence with the training content. Construct explication, and valid operationalization and measurements are clearly crucial for investigating trainability of empathy, but equally important is the assurance that the observed change or post-training empathy levels are in fact attributable to the training itself and that the findings are generalizable to other constructs. In addition, as a precondition, that the statistical analysis must have sufficient power to detect the training effect if it exists. We performed such analysis in our review.

Statistical Power. Two of the factors that contribute to random errors and consequently affect power of the statistical analysis are unreliability of measurements and small sample size. Information about these factors was available to us from the 26 quantitative studies and we included these factors as two of the criteria we used to evaluate quality of research designs. As seen in Figure 3, we used a 3-point scale for rating reliability (measured, reported, and no) and a 2-point scale for rating sample size (equal to or greater than 25 and less than 25). From our analysis, the % criterion attainment score for reliability is 63% and for sample size 46%. Slightly

more than half of the quantitative studies we reviewed either estimated or reported findings from previous studies regarding reliability of the empathy measures. Sample size in about 46% of these studies was equal or greater than 25, which, according to Edwards (1972), is a sufficient number of observations for each treatment to assure robustness of the significance test.

Other threats to statistical conclusion validity that were identified by the researchers of the quantitative studies we reviewed included restriction of range due to a ceiling effect because the participants evidenced high pre-training empathy scores, which made it difficult to measure the full range of improvement in empathy (Ancel, 2006; Beddoe & Murphy, 2004; Herbek & Yammarino, 1990), and attrition bias where some researchers reported a substantial loss of subjects during the course of the study (Beddoe & Murphy, 2004; Gantt et al., 1980).

Experimental Designs. From the experimental designs used by the 26 quantitative studies, we summarized the following information regarding the use of control group and pre- and post-testing:

- 22 (85%) of the 26 studies that used quantitative designs used either a control group (82%) or a comparison group (18%)
- Of the 22 studies that used a control group or a comparison group, eight (36%) did not randomly assign participants to treatment and control/comparison groups; 14 (64%) did.
- Of the 22 studies that used a control group or a comparison group, 11 (50%) did not check the comparability of the participants between the two groups.

- Of the 26 quantitative studies, 16 (62%) used pre- and post-tests, four (15%) used post-test only, four (15%) used pre- and post-tests plus follow-up, and two (8%) used post only and follow-up.
- The sample sizes for the 26 quantitative studies ranged from 16 to 300. Of the 26 quantitative studies, 8 (31%) used volunteers as their participants.

Based on our design quality-rating scheme, the 26 quantitative studies received percent criterion attainment scores of 63% and 73% for use of control group and use of pre-training measure respectively. It seems that the most common experimental design is the randomized experimental design with pre- and post-testing. This is a strong experimental design and can produce valid training impact estimates if implemented correctly. However, we have very limited information about implementation of the experimental design. Two researchers reported cross condition contamination because the control and the experimental groups were in regular contact with each other, which compromised the group contrast (Haynes & Avery, 1979; Shapiro et al., 2006). One study reported non-parallel pre- and post-tests (Barone et al., 2005), and one study reported multiple-testing effect (Seto et al., 2006).

External Validity. A strategy for determining external validity of findings with a single study is to examine the moderator variable effects (Shaddish, Cook & Campbell, 2002). In our review, we found that 34 % of the 29 quantitative and qualitative studies we reviewed either empirically analyzed or conjectured through logical reasoning the effects of potential moderator variables on the generalizability of findings across constructs. The external validity threats identified by the researchers include heterogeneous or non-representative samples (that is, participants' background either varied widely on such demographics as age, education, number of children, relationship status, and prior experience, e.g.,Fernandez-Olano et al., 2008; Herbek & Yammarino, 1990; Long et al., 2006; McConnell & LeCapitaine, 1988; Seto et al., 2006) or restricted samples consisting of women only (Layton, 1979; Dasgupta & Charon, 2004) or with only one male (Herbek & Yammarino, 1990).

Summary. In Figure 9, we present a summary of our ratings of the 14 design criteria.
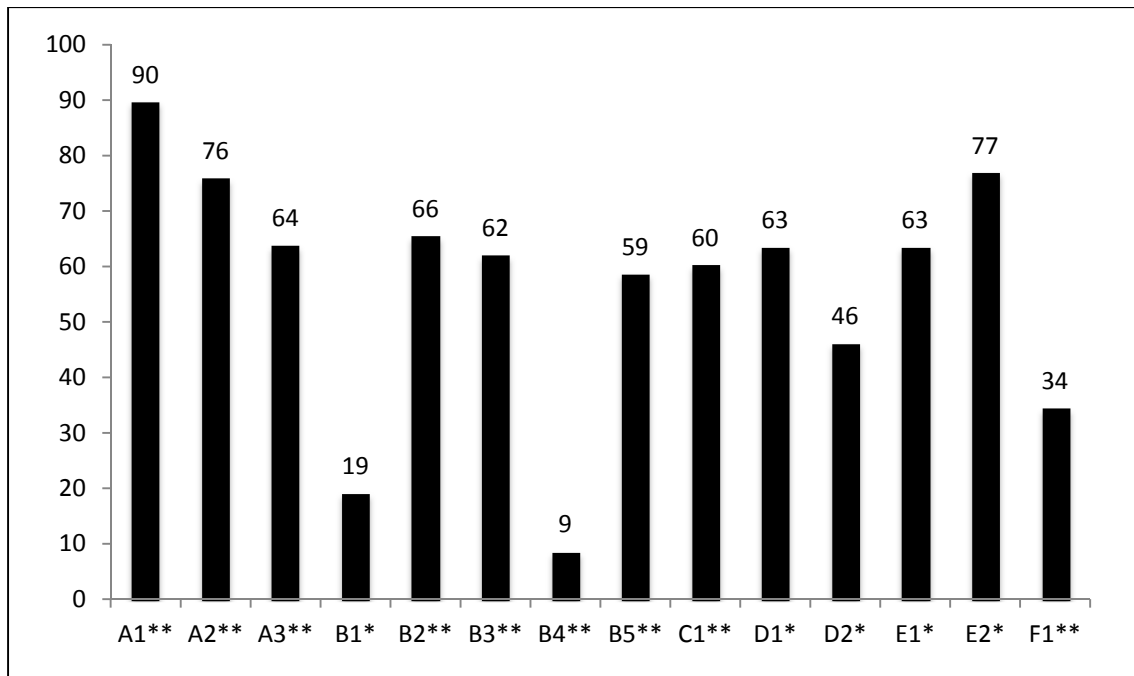
Figure 9. Average Score Attained on Each of the Design Criteria

Legend
A1** Training method definition stated.
A2** Description of training provided.
A3** Congruence between training method definition and training implementation description.
B1* Validity.
B2** Definition of empathy provided.
B3** Congruence between measurement content and training content.
B4** Congruence between empathy definition, training content and measurement content.
B5** Highest level of outcomes measured (Kirkpatrick's outcome levels).
C1** Examination of representativeness of trainee sample.
D1* Reliability.
D2* Sample size.
E1* Use of control group.
E2* Use of pre-training measure.
F1** Study checked for moderator variables.

*Design criteria that applied only to the 26 quantitative studies.
**Design criteria that apply all 29 qualitative and quantitative studies.

As seen in Figure 9, the design features employed most frequently or fully by the 29 studies were defining and describing the training, and, among the 26 quantitative studies, the use of pre-training measures. The weakest design features relate to validity of the empathy measures. The 29 studies, as a group, appear to be inconsistent in how they defined, provided training, and measured empathy. Although only 34% of these studies examined moderator variable effects on their findings, this practice related to external validity is not as critical as that for the three types of validity. Also, although 46% of the 26 quantitative

studies had sample size of 25 or larger, three studies had sample sizes of 23 and 24, which were quite close to our cutoff size of 25. On the other hand, researchers in two studies did not report their sample sizes at all. For most of all the other criteria, the % attainment scores were in the 60% range, more than half of the studies we reviewed included those design features in part or in full in their studies.

Collectively, the quality of experimental designs and the number of methodological flaws identified by the researchers seem typical of social science research and do not appear severe enough to render the training impact conclusion drawn from the 26 quantitative studies invalid. It seems quite respectable that researchers in one third of the 29 studies concerned themselves with external validity. However, together with the aforementioned issues pertaining to construct validity, specifically conceptualization and measurement of empathy and incongruence between measurement and training of empathy, we surmise that findings about empathy trainability must be interpreted carefully and accepted cautiously. We cannot effectively train people to improve on an attribute without first defining the attribute precisely and measuring it validly.

Conclusion and Discussion

In this last section of our paper, we would like to comment on the limitations of our review, and provide some recommendations for future research on the trainability of empathy in view of the findings from our review regarding evaluation practices in this area of research.

Limitations of Our Review

Our decision to use the narrative review procedure instead of meta-analysis was based on a number of factors. First, in reading the selected articles, we noticed that the authors conceptualized empathy rather differently, which therefore made aggregation of data across studies as performed in meta-analysis inappropriate. We were also aware that findings of moderating effects were reported from a good portion of the studies we reviewed, and, narrative review does "allow examination of potential moderators of the generalizability of intervention effects" (Shaddish, Cook & Campbell, p.422). Narrative review also enables researchers to combine experimental and non-experimental research findings. Although our pool of studies did not include non-experimental research, we did have some qualitative studies and we were able to include them in a majority of our analyses. Finally, the focus of our study is not just on synthesizing findings across studies pertaining to trainability of empathy, but also on examining both the practices and methodological issues in empathy trainability research. We feel that given the context and purpose of our research, narrative review clearly serves our needs better than meta-analysis. However, there are shortcomings inherent in narrative reviews and in our research.

As noted by Shaddish, Cook and Campbell (2002), narrative reviews are not very precise because the review procedure uses box score or vote counting that does not take into consideration size of the difference between groups and the probability (p) levels of statistical significance testing. In our review, we tallied the studies that reported significant findings (based on the significance levels

chosen by these studies). They also expressed a concern for the difficulty of examining and reporting relationships among outcomes and potential moderating variables because managing and understanding moderating effects based on narrative descriptions could be overwhelming if the number of articles is large. However, that was not a concern in our study since the number of articles in our article pool is relatively small.

In our review of the quality of designs used in the 29 empathy trainability research studies, our design quality criteria did not cover all the validity threats, and we focused a lot more on construct validity than internal, statistical conclusion and external validity. The former limitation is due to the restricted amount of data available to us in the articles we reviewed. Our emphasis on construct validity is because of our interest in the empathy construct and its trainability. Nonetheless, our review of the design quality is not comprehensive and consequently our conclusion about methodological practices in empathy training research is not complete.

In our review of construct validity, we had to make inferences about the training and measurement content and the match between them. We all engaged in this process and we routinely checked each other's inferences, and we do feel that judgment error is minimal and that our general conclusions were in the right direction, as we usually concurred with each other. However, the inferences we made are nonetheless subjective judgments, and even with consistency, it is still not a guarantor of accuracy.

We want to note that our method of examining the match between definition, treatment content and measurement content is akin to the method of examining alignment among standards,

curricular and exam contents for assessing curricular validity of test scores or effectiveness of curriculum materials or standards (Beck, 2007; Resnick, Rothman, Slattery & Vranek, 2004). We found it a useful technique for examining construct validity of treatments and outcomes in research and evaluation, especially if it is refined with formal consistency and validity checks. The procedure can be extended to construct validity of persons and settings and it aligns very well with the philosophical orientation of the narrative review approach to examining quality of evaluation or research practices.

Recommendations for Empathy Training and Evaluation Research

We set out to search for answers to the umbrella question: Can we train people to become empathetic? Our investigation has led us to cautiously conclude that empathy is trainable with qualifications and caveats. The research findings that we reviewed seem to suggest that it is feasible to enhance one's knowledge about empathy and the skills to "act" empathically with a variety of training methods. Hence, technically, empathy is trainable. However, there is no sound evidence to support the conclusion that training can effectively change people's propensity to behave empathically after training, that is, transfer their learning and change their empathic behaviors in the natural environments. Furthermore, there is also a dearth of evidence to suggest that training can indeed alter people's affective empathy towards others. It seems logical to conclude that just knowing about and exhibiting empathic behavior without a visceral concern for others is not the ultimate goal in empathy

training. In that regard we may conclude that our goal to train people to become empathetic has not been met, at least on the basis of the research evidence available in the literature.

The failure to create empathetic people could be a function of the inadequacy of both the training and research methods used in the empathy trainability studies. With regard to the former, the training designs employed in the research we reviewed appear to lack the intensity and focus to affect transfer of learning. To understand what training strategy can affect empathy behavior beyond the training environment, we need to further ask the question: Does change in empathy behavior require an enhanced affective empathy or not? We would argue that empathic behaviors could be exhibited with or without the feeling component of empathy; that is, people can learn to behave empathetically whether or not they have empathetic feelings towards others. Empathy training should make a distinction between the two outcomes, behavioral change with and without affect change, because choosing one or the other will affect the design and effort of the training.

Regarding methodology used in the studies we reviewed, the major shortcomings we noticed that are unique to this line of research were: unclear and inconsistent conceptualization and operationalization of empathy, minimal evidence in support of the measurement validity, mismatch between measurement and training content, a lack of treatment implementation data, incomplete description of study design, and a general neglect of the measurement of transfer of learning. It is difficult to draw conclusions about trainability of empathy and the extent to which the findings hold over variations in persons, settings, treatments and outcomes (Shaddish, Cook, & Campbell, 2002), if researchers do not clearly define empathy, provide training as defined, measure what is being trained, and collect ample data about the research design, settings and permanent and persistent change in empathic behavior. We need sound scales to measure the feeling manifestations of empathy, and resources and techniques to measure empathetic behavior change in the social milieu. Not only do we feel that more evaluation research should be conducted to examine the trainability of empathy, these future studies should take into consideration the methodological issues identified in our review, and they should also address other more in-depth theoretical and methodological issues as exemplified by the following research questions:

- To train people to become more empathetic, do we need to focus on all three components (knowledge, affect and behavior)? In other words, does being empathetic require all of the three components? If one empathy component is enhanced through training, will the other non-trained components also be enhanced?
- What are the key bio-demographic characteristics of the target population that can benefit the most from empathy training and what are the potential interactions between these trainee characteristics and training methods?
- If affect and behavior empathy can be increased through training, is the change permanent?

We hope that our findings regarding the methodological issues in training and

evaluation research can help future researchers in their quest for knowledge about how we can help people and ourselves become more empathetic.

References

Ancel, G. (2006). Developing empathy in nurses: An in-service training program. *Archives of Psychiatric Nursing, 20*(6), 249-257.

Aspy, D. N., Roebuck, F. N., & Aspy, C. B. (1984). Tomorrow's resources are in today's classroom. *Personnel and Guidance Journal, 62*(8), 455-459.

Avery, A. W., & Thiessen, J. D. (1982). Communication skills training for divorcees. *Journal of Counseling Psychology, 29(*2), 203-205.

Barak, A., Engle, C., Katzir, L., & Fisher, W. A. (1987). Increasing the level of empathic understanding by means of a game. *Simulation and Games, 18*(4)*, 458–470.

Barone, D. F., Hutchings, P. S., Kimmel, H., Traub, H. L., Cooper, J. I., & Marshall, C. M. (2005). Increasing empathic accuracy through practice and feedback in a clinical interviewing course. *Journal of Social and Clinical Psychology, 24*(2), 156-171.

Barrett-Lennard, G.T. (1962). Dimensions of therapist response as causal factors in therapeutic change. *Psychological Monographs, 76*, 43 (Whole No. 562).

Baucum, D.H. (1985). A review of the California psychological inventory. In James V. Mitchell (Ed.). *Ninth Mental Measurement Yearbook.* Lincoln, Nebraska: The University of Nebraska Press.

Beck, M.D. (2007). Review and other views: "Alignment" as a psychometric issue. *Applied Measurement in Education, 20*(1), 127-135.

Beddoe, A. E., & Murphy, S. O. (2004). Does mindfulness decrease stress and foster empathy among nursing students? *Journal of Nursing Education, 43*(7), 305-312.

Block-Lerner, J., Adair, C., Plumb, J. C., Rhatigan, D. L., & Orsillo, S. M. (2007). The case for mindfulness-based approaches in the cultivation of empathy: Does nonjudgmental, present-moment awareness increase capacity for perspective-taking and empathic concern? *Journal of Marital and Family Therapy, 33*(4), 501-516.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5*(4), 437-474.

Carkhuff, R. R. (1969). *Helping and human relations.* New York: Holt, Rinehart & Winston.

Crabb, W. T., Moracco, J. C., & Bender, R. C. (1983). A comparative study of empathy training with programmed instruction for lay helpers. *Journal of Counseling Psychology, 30*(2), 221-226.

Dasgupta, S., & Charon, R. (2004). Personal illness narratives: Using reflective writing to teach empathy. *Academic Medicine, 79*(4), 351 – 356.

Davis, M. H. (1980). Individual differences in empathy: A multidimensional approach. *Dissertation Abstracts International, 40* (7-B), 3480.

Davis, M., & Franzoi, S. (1991) Stability and change in adolescent self-consciousness and empathy. *Journal of Research in Personality, 25*, 70-87.

Dymond, R. F. (1949). A scale for the measurement of empathic ability. *Journal of Consulting Psychology, 13*, 127-133.

Edwards, A.L. (1972). *Experimental designs in psychological research

(Fourth Edition). New York: Holt, Rinehart & Winston, Inc.

Evans, B. J., Stanley, R. O., & Burrows, G. D. (1993). Measuring medical students' empathy skills. *British Journal of Medical Psychology, 66*(2), 121-133.

Feighny, K. M., Monaco, M., & Arnold, L. (1995). Empathy training to improve physician–patient communication skills. *Academic Medicine, 70*(5), 435–436.

Fernández-Olano, C., Montoya-Fernández, J., & Salinas-Sánchez, A. S. (2008). Impact of clinical interview training on the empathy level of medical students and medical residents. *Medical Teacher, 30*(3), 322-324.

Fraser, J. A. H., & Vitro, F. T. (1975). The effects of empathy-training on the empathic response levels and self-concepts of students in a teacher-training program. *Canadian Counsellor, 10*(1), 25-28.

Gantt, S., Billingsley, D., & Giordano, J. A. (1980). Paraprofessional skill: Maintenance of empathic sensitivity after training. *Journal of Counseling Psychology, 27*(4), 374-379.

Gurman, A. S., & Kniskern, D. P. (1978). Research on marital and family therapy: Progress, perspective, and prospect. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (2 ed.). New York: Wiley.Hatcher, S. L., Nadeau, M. S., Walsh, L. K., Reynolds, M., Galea, J., & Marz, K. (1994). The teaching of empathy for high school and college students: Testing Rogerian methods with the Interpersonal Reactivity Index. *Adolescence, 29*(116), 961-974.

Haynes, L. A., & Avery, A. W. (1979). Training adolescents in self-disclosure

and empathy skills. *Journal of Counseling Psychology, 26*(6), 526-530.

Herbek, T. A., & Yammarino, F. J. (1990). Empathy training for hospital staff nurses. *Group and Organization Management, 15*(3), 279-295.

Higgins, E., Moracco, J., & Danford, D. (1981). Effects of human relations training on education students. *Journal of Educational Research, 75*(1), 22-25.

Holm, U. (1996). The affect reading scale: A method of measuring prerequisites for empathy. *Scandinavian Journal of Educational Research, 40*(3), 239-253.

Jackson, E. (1985). The relationship of peer rated warmth, understanding, and openness to affective sensitivity: A multivariate analysis. *Psychology: A Quarterly Journal of Human Behavior, 22*(3), 28-55.

Kagan, N., & Schneider, J. (1987). Towards the measurement of affective sensitivity. *Journal of Counselling and Development.* 65 (9), 459-464.

Kirkpatrick, D.L. (1994). *Evaluating training programs: The four levels.* San Francisco, CA: Berrett-Koehler.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development.* Upper Saddle River, New Jersey: Prentice Hall.

Layton, J. M. (1979). The use of modeling to teach empathy to nursing students. *Research in Nursing and Health, 2*(4), 163-176.

LaMonica, E.L., Carew, D.K., Winder, A.E., Haase, A.M.B & Blanchard, K.H. (1976). Empathy training as a major thrust of a staff development program. *Nursing Research, 25*(6), 447-451.

Long, E. C. J., Angera, J. J., Carter, S. J., Nakamoto, M., & Kalso, M. (1999). Understanding the one you love: A

longitudinal assessment of an empathy training program for couples in romantic relationships. *Family Relations, 48*(3), 235-242.

Long, E. C. J., Angera, J. J., & Hakoyama, M. (2006). Using videotape feedback during intervention with married couples: A qualitative assessment. *Family Relations, 55*(4), 428-438.

McConnell, D. M., & LeCapitaine, J.E. (1988). The effects of group creativity training on teachers' empathy and interactions with students. *Reading Improvement, 25*(4), 269-275.

Mehrabian, A. & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality, 40*, 525-543.

Nerdrum, P. (1997). Maintenance of effect of training in communication skills. A controlled follow-up study of level of communicated empathy. *British Journal of Social Work, 27*(5), 705-722.

Nerdrum, P., & Ronnestad, M. H. (2002). The trainees' perspective: A qualitative study of learning empathetic communication in Norway. *The Counseling Psychologist, 30*(4), 609-629.

Resnick, L.B., Rothman, R., Slattery, J.B., & Vranek, J.L (2004). Benchmarking and alignment of standards and testing. *Educational Assessment, 9*(1-2), 1-27.

Ridley, C. A., Jorgensen, S. R., Morgan, A. G., & Avery, A. W. (1982). Relationship enhancement with premarital couples: An assessment of effects on relationship quality. *American Journal of Family Therapy, 10*(3), 41-48.

Sechrest, L., West,S.G., Phillips,M.A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatments. In L. Sechrest, S.G. West, M.A. Phillips, R. Redner, & W. Yeaton. (Eds.) *Evaluation Studies Review Annual* (Vol.4), pp.15-35. Thousand Oaks, CA: Sage.Salas, E. & Cannon-Bowers, J. A. (2001). The science of training: A decade of Progress. *Annual Review of Psychology, 52*, 471-499.

Seto, A., Young, S., Becker, K. W., & Kiselica, M. S. (2006). Application of the triad training model in a multicultural counseling course. *Counselor Education and Supervision, 45*(4), 304-318.

Shadish, W.R., Cook T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Shapiro, J., Rucker, L., Boker J., & Lie, D. (2006). Point-of-View writing: A method for increasing medical students' empathy, identification and expression of emotion, and insight. *Education for Health, 19*(1), 96-105.

Shapiro, S. L., Schwartz, G. E., & Bonner, G. (1998). Effects of mindfulness-based stress reduction on medical and premedical students. *Journal of Behavioral Medicine, 21*(6), 581-599.

Summerfelt, W.T. (2003). Program strength and fidelity in evaluation. *Applied Development Science, 7,* 55-61.

Warner, R.E. (1984). Enhancing teacher affective sensitivity by a videotape program. *Journal of Educational Research, 77*(6), 366-368.