Do Online Matchmaking Tests Work? An Assessment of Preliminary Evidence for a Publicized 'Predictive Model of Marital Success'

James Houran
TRUE, LLC

Rense Lange
Integrated Knowledge Systems, Inc.
Illinois State Board of Education

P. Jason Rentfrow and Karin H. Bruckner
TRUE, LLC

Acceptance of online matchmaking as a culturally legitimate approach to mate selection - and consumer spending on these services - continues to rise. However, online matchmakers' escalating claims that their services derive from scientific methods remain questionable because solid empirical evidence for such claims is rarely offered. Unfortunately, even when available, the quality of such evidence leaves much to be desired due to conceptual as well as technical problems. Such issues are illustrated here by a detailed analysis of an instructive case study of an attempt to validate aspects of the commercial eHarmony.com dating service. Apart from identifying serious logical flaws that invalidate the case study's conclusions, additional shortcomings are identified related to the involved variables, research design, and sampling biases. Because such issues almost certainly play a role in online dating and related research, the paper concludes with a discussion of modern test construction approaches derived from Item Response Theory, and Rasch scaling in particular, that can be used to identify and sometimes correct many of the problems described here. Online dating services must solve many of the problems outlined here to remain a viable and acceptable area of practice and research.

Compatibility seems to be a new buzz word in the vernacular of Western popular culture, considering the wealth of media coverage of the rising trend for online dating companies to conduct romantic matchmaking based on stated personal preferences and alleged

Author info: Correspondence should be addressed to James Houran, TRUE, LLC, Central Tower at Williams Square, Suite 1600, 5215 N. O'Connor Blvd, Irving, TX, 75039; e-mail: jim.houran@true.com.
North American Journal of Psychology, 2004, Vol. 6, No. 3, 507-526.

personality testing (Egan, 2003; Goot, 2004; Mand, 2004; Mulrine, 2003). This trend is highly relevant to the field of personal relationships since Internet dating services represent a significant and growing segment of online services and the general personals and dating services. Market data for last year alone reveals that web services accounted for approximately 43 percent of the $991 million United States dating-service sector, which also includes print and radio personal ads and other offline operations.[i] Consumers tripled their spending on Internet dating services between 2001 and 2002, and Jupiter Research expects online dating sites to record over $640 million by 2007. Some have estimated that as many as 22 percent of the 98 million singles in the U.S. in 2002 used online dating.[ii] As the industry segment grows, its advertising is becoming ubiquitous. Between 2000 and 2003, the number of online advertisements for internet dating services increased six-fold.[iii] As the stigma historically associated with Internet dating is seemingly diminishing, these services are targeting and reaching their intended audiences with unprecedented success. According to Hitwise, visits to online dating sites have increased by 45 percent since January 2003, with the online dating sector now representing 0.8 percent of all online visits.[iv]

More individuals are using online compatibility testing for mate selection, so it is crucial for our field to become more discerning about the scientific validity of these matching methods. Houran (2004) recently discussed in an address to the online dating industry that some companies and their claims of scientific validity for their 'compatibility tests' are difficult to evaluate because psychometric data are either not collected or rarely made available for public scrutiny – including peer review in the scientific literature. Finn and Banach (2000) similarly noted the difficulties of ascertaining the credentials and identity of service providers, accessing accurate information, reliance on untested methods, difficulties in online assessment, and the lack of standards and regulation regarding online human service practices. This opinion was also echoed and expanded in a high profile article recently published in the *American Psychologist* (Naglieri, Drasgow, Schmit, Handler, Prifitera, Margolis, & Velasquez, 2004).

While many services advertise that their matching methods are "scientifically based" and are capable of identifying a better match for a member than other services, Houran's (2004) review revealed that virtually none of these services provide acceptable substantiation for their claims. For example, eHarmony.com claims, among other things, that its "patented matching technology" is "based on 35 years of empirical and clinical research on what goes into successful relationships, and it brings together singles using a scientifically-proven set of compatibility principles" that involves "29 dimensions that create compatibility and

play a key role in the most successful relationships." Their web site further explains:

> eHarmony's service is underpinned by its highly accurate, patented scientific model for matching. It is squarely built upon research conducted with more than 5,000 married persons. From careful statistical analyses of this data, a team of Ph.D. psychologists led by eHarmony founder Dr. Neil Clark Warren extrapolated a series of insights and understandings about relationships.[v]

This specific study seems to be the foundation of eHarmony's tests and services, but it is not referenced in detail and a copy of the full analyses and results are neither posted for customers nor otherwise offered to the public for evaluation. Moreover, Neil Warren Clark was not listed as an author when validity data on eHarmony were recently presented to a professional audience. These data will be discussed at length below.

Match.com and PerfectMatch.com make similar claims of scientific support. Match.com states on its site that its "Personality Matching" method of pairing potential mates was "developed in partnership with weAttract.com and [is] based on a 15-year research initiative by weAttract scientists and inventors Drs. Mark Thompson and Glenn Hutchinson . . . ."[vi] It offers no further substantiation or explanation.

PerfectMatch.com goes even further in its claims than does Match.com. On its website, PerfectMatch claims that objective science materially distinguishes it from its competitors:

> Unlike online dating sites, our approach utilizes a scientifically based personality profiler developed by leading relationship expert, Dr. Pepper Schwartz, Ph.D. Our more refined approach, based on over 30 years of research, results in fewer, but more qualified matches. Dr. Schwartz holds the distinction of being the only relationship expert on the Web who's a published authority, as well as a professor at a major U.S. university.[vii]

Rather than citing empirical evidence in support of its methods, PerfectMatch provides a list of positive endorsements of its methods from academicians with university affiliations. Unfortunately, it is not clear from these testimonials whether these academicians actually have expertise in the specific field of modern test construction and validation. The site also draws an explicit link between its methods and a more widely recognized personality test, stating that "[o]ur test is based on the same theory that inspired the famous Myers Briggs Type Indicator®. We

strongly believe it's much more sophisticated and effective than any romantic matching tool on or off the net."[viii]

The only patented online matching system or compatibility test belongs to eHarmony (U.S. Patent No. 6,735,568). Although reliability and validity data for its "method and system for identifying people who are likely to have a successful relationship" are not included in the patent document, eHarmony did present preliminary validity data at the 16th Annual Convention of the American Psychological Society (Carter & Snow, 2004). In addition, the full paper (termed a 'handout' by the authors) corresponding to this presentation was recently posted by eHarmony (see: http://static.eharmony.com/images/eHarmony-APS-handout.pdf). This is a clear and positive step towards satisfying the concerns of Houran (2004), Finn and Banach (2000), and Naglieri et al. (2004).

The web posting of Carter and Snow's work further permits an initial assessment of eHarmony's public claims that its matching methods have scientific validity. While we share Carter and Snow's (2004) belief that it is necessary to examine the validity of such matching systems, we must, unfortunately, take issue with their research design and statistical techniques. To the extent allowed for by the available information, we first present an overview of eHarmony's overall approach and then offer criticisms and suggestions for how issues of validity could be addressed in superior ways.

eHarmony Testing Background and Theoretical Rationale

Carter and Snow (2004) did not describe the nature of eHarmony's matching approach, yet glimpses into the rationale behind this service are provided by this company's patent application (Buckwater, Carter, Forgatch, Parsons, & Warren, 2004). This application states that couples' relationship satisfaction is predicted from a large set of predictive background variables that are reduced to a smaller set via factor analysis. Next, "the results are studied to identify the candidate and user combinations that would result in the most satisfaction" (p. 4). This is achieved by "approximating a relationship between the individual satisfaction index … [which] … includes, but is not limited to, performing a multiple linear regression and correlation analysis on the individual satisfaction indexes versus the factor data" (p. 8). The patent application specifies that Spanier's (1976; Spanier & Cole, 1976) *Dyadic Adjustment Scale* (DAS) was used as the criterion variable in these analyses.

In other words, eHarmony combines its predictor variables in order to maximize couples' DAS scores. It is important to note that the DAS was also used by Carter and Snow (2004). This frequently used scale

conceptualizes marital quality not only as a subjective evaluation but also a process in a dyad. The DAS has a total score, as well as four subscales that measure Dyadic Cohesion ("Do you and your mate engage in outside interests together?"), Dyadic Consensus (the extent of agreement/ disagreement between the couple on various issues), Dyadic Affection ("Do you kiss your mate?"), and Dyadic Satisfaction ("How often do you discuss or have you considered divorce, separation or terminating your relationship?").

Spanier constructed the DAS via classical test theory, which unfortunately does not guarantee that this instrument meets the requirements of modern approaches - such as Rasch scaling (see e.g., Bond & Fox, 2001) - which can yield interval-level measures free of response biases related to extraneous variables, such as a respondent's age and gender. As will be discussed in more detail in a later section, it is important to control for such biases because statistical theory (Stout, 1987) and computer simulations alike (Lange, Irwin, & Houran, 2000) indicate that response biases can lead to spurious factor structures, significant distortions in scores, and consequently erroneous research findings. In this light it is not surprising to observe that the DAS' factor structure, and hence its construct validity, continues to be debated in the literature (e.g., Hunsely, Pinsent, Lefewbvre, James-Tanner, & Vito, 1995; Sharpley & Cross, 1982; Spanier & Thompson, 1982).

On a broader level, eHarmony's website and advertisements stress that its theoretical orientation is based on the notion that romantic compatibility equates with greater *similarity* than dissimilarity between two individuals. To be sure, many studies indicate that individuals are more likely to select marital partners (or be happier in their relationships) with similar, as opposed to dissimilar, personality characteristics (for reviews and discussions see e.g., Gottman, Murray, Swanson, Tyson, & Swanson, 2002; Holman, 2001; Ickes, 1985; Karney & Bradbury, 1995). Comparable findings extend to couples' perceived similarity across the variables of physical and social attractiveness, socioeconomic status, and level of intelligence (Sussman & Reardon, 1987). This hypothesis of matching or homogamy (the mating of similar individuals) is not limited to traditionally married partners, but also applies to heterosexual cohabiting and homosexual couples (Kurdek & Schmitt, 1987).

However, although a thorough evaluation of the theoretical orientation of eHarmony's (or any similar company) test is not an explicit focus of this paper, it is important to note the limitations of a strict homogamy approach insofar as it could affect the relationship quality of paired individuals in research studies. Specifically, other research underscores the notion of *complementarity* (e.g., Houts, Huston, & Robins, 1996), which entails that couples achieve compatibility by

harmonizing, or even exploiting, differences in partners' interpersonal styles and life skills. For instance, Dryer and Horowitz (1997) found that individuals in complementary partnerships (submissive people with dominant partners, dominant people with submissive partners) reported more satisfaction than did those with partners whose styles were similar. Moreover, individuals complementing their partner's behavior were more satisfied with couple interactions than were those whose goals were not complementary.

Aronson, Wilson, and Akert (1994, p. 386) concluded that the research evidence for complementarity is mixed at best, and based on a few studies only. By contrast, the findings of Dryer and Horowitz (1997) and related literature (e.g., Beach, Whitaker, Jones, & Tesser, 2001; Bor, Prior, & Miller, 1990; Carroll, Gilroy, Hoenigmann-Stovall, & Turner, 1998; Gross & McIlveen, 1998; Kerckhoff & Davis, 1962; Lange, Jerabek, & Houran, 2004; Nowicki & Yaughn, 1999; Pilkington, Tesser, & Stephens, 1991) all increasingly speak to the importance of complementarities across various relationship types. This seems consistent with the four general perspectives on compatibility as delineated by Levinger (1986). These perspectives include: (1) the relationship among the partners' values, personalities, and predispositions, (2) the patterns of accommodation adopted by a couple, (3) the couple's adaptability to each other's needs in the face of mutual conflict (mutual transformation), and (4) temporal changes or convergences in preferences, goals, and dispositions (dispositional transformation). Thus, broadly speaking, complementarity implies that successful couples integrate qualitatively different issues into the relationship.

Of course, some authors claim that dating online can be quite different than dating offline given the nature of cyberspace (Whitty, 2003; Whitty & Carr, 2003). On the other hand, other research seems to confirm that basic compatibility issues with online dating parallel those with offline dating (e.g, Baker, 2002). Clearly, additional research is needed in order to better understand whether the variables that make a successful relationship offline also make a successful relationship with someone first met online.

Summary of Carter and Snow (2004)

In their study on the predictive models of marital success, Carter and Snow (2004) attempted to show that married individuals who were paired using eHarmony's matching system are more satisfied in their relationships than individuals who were not paired by any particular matching system. To this end, they gathered data on relationship satisfaction from a sample of married individuals registered at

eHarmony's website as well as an independent sample of married individuals. They then compared the two groups of married persons on the satisfaction measure and concluded that those individuals who met via eHarmony's matching system were significantly more satisfied in their relationship than those individuals who met offline.

On the surface, this experimental design appears to be a valid way for determining the efficacy of eHarmony's matching system. However, as we discuss below, there are several shortcomings associated with the approach, which invalidate any of the conclusions one could draw about eHarmony's matching system from these data.

Conceptual Issues

As we already noted, Carter and Snow's (2004) eHarmony respondents and the control respondents both completed the DAS. The authors' main argument for the validity of eHarmony's methods derives from the fact that the eHarmony respondents obtained higher DAS scores than did the control group. However, by definition, the eHarmony couples exist *precisely* because their members' predicted DAS score would be optimal. As a result, two people that were matched by eHarmony *by construction* should have higher DAS scores than would randomly selected couples. Accordingly, the greater DAS scores observed for the eHarmony couples can be explained as an artifact that merely reflects eHarmony's success in predicting DAS scores from the available background variables. Carter and Snow's (2004) findings thus mainly illustrate that couples that were *a priori* selected to have higher DAS scores do indeed possess such higher scores. Unfortunately, we would argue that we do not know what higher DAS scores mean exactly, given the fact that this measure has not been shown to meet modern methods of test construction and validation.

Of course, the real question to be answered in this context is not whether DAS scores can be predicted, but whether couples with higher DAS scores indeed have better relationships or marriages. In this respect Carter and Snow's (2004) research provides no definitive answers, because it does not meet the absolutely critical criteria that the groups being compared for this purpose must be similar on all key variables in order to reach meaningful conclusions. This was not the case in Carter and Snow's (2004) investigation, because the comparison group differed significantly from the eHarmony group in terms of *Age* (eHarmony users were older), *Education* (eHarmony users had more education), *Income* (eHarmony users had higher income), *Number of Years Dated Before Marrying* (eHarmony users dated less), *Number of Years Married* (eHarmony users were married for a shorter length of time), and the *Number of Previous Marriages* (eHarmony users had more marriages).

To adjust for these gross sample differences, Carter and Snow (2004) applied Analyses of Covariance (ANCOVA) on DAS scores and thus hoped to control for the variables listed above. We address later that these types of methods are ineffective for countering such confounding sources of bias. But even apart from these issues, the findings are ambiguous as they contradict eHarmony's assumption that more similar couples are also more compatible. For instance, even when Age, Education, Income, Years of Dating, Years Married, and Number of Previous Marriages are taken into account, the eHarmony couples showed *greater* diversity in DAS scores than did the control couples (see Table 2: Carter & Snow, 2004, p. 4; available online at: http://static.eharmony.com/images/eHarmony-APS-handout.pdf). In particular, the eHarmony group showed considerably *greater* variance (S^2 = 525.53) on the DAS total score than did the control group (S^2 = 336.81).[ix] As is indicated by the parenthesized values, the eHarmony group likewise showed greater variance on the Dyadic Consensus (S^2 = 125.77 vs. 78.29), Dyadic Satisfaction (S^2 = 73.55 vs. 47.36), Affectional Satisfaction (S^2 = 9.82 vs. 6.87), and Dyadic Cohesion (S^2 = 24.48 vs. 15.47) subscales of the DAS. Thus, the eHarmony couples' scores on the DAS thus appear more *dissimilar* than those of the control couples. Accordingly, Carter and Snow's data support a complementarity view of romantic compatibility, rather than a strict homogamy view of compatibility.

Technical Aspects

In addition to the above, we note the following technical shortcomings.

Prediction of Relationship Success. Carter and Snow (2004) sampled newlyweds who found their spouses using eHarmony's services. This approach is problematic because it assumes that marriage in itself is an inherent indicator of relationship success. This assumption is particularly questionable given that over 50% of all marriages end in divorce (e.g., Bramlett & Mosher, 2002), and that most marriages dissolve within the first 5 years. Accordingly, studying relationship stability and satisfaction simultaneously across multiple types of relationships – e.g., marriage, cohabitation, homosexual unions, long-term dating – seems to us to be a more comprehensive approach for isolating and validating those variables that underpin successful romantic relationships.

Research Design. Carter and Snow (2004) used a cross-sectional approach in which they compared responses from the eHarmony sample to an independent sample of married persons. Although this approach has some advantages (i.e., it is practical, efficient, and inexpensive), it cannot fully test the efficacy of any matching system. Such designs only provide

a snapshot of existing differences in the attitudes, beliefs, and behaviors of the groups. However, cross-sectional designs do not shed light on the origins of couples' similarities or differences. In Carter and Snow's (2004) case, any observed difference might well reflect the result of diversity in the demographic characteristics of the samples, and not necessarily in the effectiveness of the matching system. Indeed, cross-sectional designs that compare samples that are grossly different in terms of age, race, education, and number of years married can only reveal differences in psychological characteristics as a function of the socio-cultural characteristics that define them. Thus, cross-sectional methodologies can be very informative, but only in so far as the groups being compared are similar on key demographics.

A more thorough, albeit time consuming and expensive, methodology for evaluating the effectiveness of matching systems on relationship success is a longitudinal methodology that assesses participants' attitudes, beliefs, and behaviors at various points in time for a period of months or years. Such designs not only shed light on differences between groups, but more importantly, they have the power to inform our understanding of how those differences develop and evolve over time. For instance, previous research (Gottman, 1994; Huston, Caughlin, Houts, Smith, & George, 2001; Huston, Niehuis, & Smith, 2001) shows that newlyweds are significantly more intimate and committed in their relationships than are couples that have been married for 3 years or more. Therefore, longitudinal research could show whether (1) newlyweds paired through a matching system display levels of intimacy and commitment similar to newlyweds paired through idiosyncratic methods, and (2) couples who were paired by a matching system display a similar marital satisfaction and commitment trajectory over time as couples paired through idiosyncratic methods. For instance, it could be that the high levels of intimacy common among all newlyweds hold over time for couples paired through a matching system, but not for couples paired through idiosyncratic methods. Clearly then, to develop a more thorough and comprehensive understanding of the efficacy of any matching system on couples' satisfaction and commitment, a longitudinal design is ideal as a complement to cross-sectional research.

3. *Potential Sample Biases*. In addition to the large demographic differences in Carter and Snow's two samples, the participants in these samples were selected quite differently. We speculate that the eHarmony sample was comprised of individuals who were very motivated to forge long-term relationships, since they signed up for an online matchmaking service that catered specifically to this outcome for a fee. Such *a priori* motivations might well mean that the eHarmony participants had artificially higher levels of Sternberg's cognitive Decision/Commitment

component of his Triangular Theory of Love and Attachment (1986). According to Sternberg, the amount of love or relationship satisfaction that a person experiences is due to the strength and interaction of three components: Intimacy (the feeling of closeness and bondedness), Passion (the drives that produce romance, physical attraction, and sexual intercourse), and Decision/Commitment (the decision that one loves another and the commitment to continue that relationship).

This is an important consideration, because recent work by Lange, Jerabek et al. (2004) found that an individual's perception of his/her compatibility to a partner becomes significantly biased both quantitatively and qualitatively, as relationship satisfaction fluctuates. This finding is consistent with previous work (Cobb, Davila, & Bradbury, 2001; Levinger, 1986; Neff & Karney, 2003) that suggests global relationship satisfaction derives from the tendency to view positive perceptions as more important than negative perceptions, as well as the tendency to alter the importance of specific perceptions as needed. For example, the tendency to describe the marital relationship in unrealistically positive terms is called marital conventionalization. Such positive cognitive distortions in marriage– what Edmonds (1967) viewed as social desirability bias in marital quality measurements—are strikingly similar to psychological constructs such as positive illusions (Taylor & Brown, 1988) and unrealistic optimism (Scheier & Carver, 1992).

Thus, the assessment or cognitive appraisal of one's partner and the quality of marriage may well parallel a self-fulfilling prophecy (Houran & Lange, 2004) whose contents form a cognitive set strongly influenced by Sternberg's (1986) component of Decision or Commitment related to the relationship. Although very little information was provided about exactly how Carter and Snow obtained the control sample, it seems likely that participants were financially compensated for their participation. In particular, Carter and Snow described the recruitment of the control group as "…using a commercial online research firm from an existing panel of individuals interested in online research" (p. 2). Thus, it is likely that the difference in motivations between self-selected participants and recruited participants is a significant confounding factor. Accordingly, the differences that Carter and Snow (2004) purport as evidence for the efficacy of eHarmony's matching system probably also reflect these qualitative differences of their samples.

One solution for avoiding the sampling biases described above would be to employ a double-blind methodology. A double-blind methodology essentially ensures that neither the participants nor the experimenters know whether the participants were paired using the matching system or by some other means. For instance, half of participants could be randomly assigned to a compatibility group (i.e., a group where all

participants are matched to people using a matching system) and the other half to an incompatible group (i.e., a group where all participants are matched to people who they are not compatible with). Participants could then interact with the partner they were assigned to be with and report their level of satisfaction with and attraction to that person. If participants paired to people with whom testing indicated that they should be theoretically compatible indeed report more satisfaction and attraction than participants who were matched to people with whom testing indicated they should be theoretically incompatible, then it would be reasonable to conclude that the matching system has preliminary validity.

Analytical Technique

Another issue to consider when assessing the validity of various online matching systems pertains to measurement. In particular, the psychometric properties of the independent variables used to determine which individuals are "compatible" and which individuals are "incompatible," as well as the psychometric properties of the dependent measures used to determine whether the matching systems are indeed effective, are critical for any scientific evaluation.

Kline (1986) noted that researchers traditionally construct and validate their assessment instruments via classical test theory as embodied for instance in the use of factor analysis and reliance on the KR-20 or Coefficient Alpha as the major index of the scale's quality. Next, the sum of the scores received on the test items is taken to be a valid index of the latent trait (e.g., marital satisfaction) under consideration. Unfortunately, such techniques essentially treat all items as equivalent and ignore the possibility that some items may be more diagnostic of individuals exceptionally high in the particular construct than are other items (Bond & Fox, 2001). Another major flaw of the raw score approach is that summed scores do *not* provide linear (i.e., interval-level) measures of the underlying trait – moreover, the approach does not recognize that some items may be biased such that subjects with *identical trait levels* receive systematically *different* scores. This might be the case, for instance, when women (or younger respondents) endorse some questions more (or less) often then do men (or older respondents) with *equal* trait levels. Thus, traditional scaling approaches offer no indication of the true internal validity of respondents' scores.

Carter and Snow (2004) neither reported on the psychometric properties of their dependent measures, nor did they describe the methods that were used to examine the psychometrics of the statistical matching system used by eHarmony. Therefore, we are left to assume that their methods and measures are reliable, at least based on classical testing

standards (i.e., coefficient alphas ≥ .70, see Kline, 1986). Yet, it remains unknown whether any of their instruments actually meet even these lax and partially-developed testing standards, much less the more stringent standards adopted by modern theories of psychological measurements (cf. American Educational Research Association, American Psychological Association, & National Council on Measurement, 2002). Methods for addressing these standards are included in our concluding remarks.

DISCUSSION

Interestingly, Internet companies like eHarmony, Match, and PerfectMatch did not pioneer computer matchmaking. In 1956 Art Linkletter, host of the popular television show *People Are Funny*, matched a couple using a computer. *Time* magazine (Nov 19, 1956) reported, "Remington Rand's Univac No. 21 turned Cupid, brought together a flesh-and-blood couple as scientifically selected 'ideal marriage mates'" (p. 79). *Time* also reported that the couple was paired based on a 32-item questionnaire developed by "The Father of American Marriage Counseling," Paul Popenoe. The happy couple became engaged, and Art Linkletter offered to pay the airfare for their Paris honeymoon. Following the Univac No. 21 experiment, the computer dating craze blossomed through the 1970s and 80s.

The recent advent of the Internet and a plethora of online dating services have now radically expanded opportunities for singles to pursue relationships via computerized matchmaking (Ahuvia & Adelman, 1992; Whitty, 2003) – to be sure, one only needs to browse a selection of these websites to see collections of "testimonials" from couples who met through these services and are now married. In this sense, computer matchmaking has evolved from an entertainment vehicle to a commercial enterprise that is often being advertised to the public as a health and human service operated by relationship and testing experts. Disturbingly, scientific evidence for such matching systems is not forthcoming even when the public and researchers request it (Houran, 2004). Carter and Snow's (2004) research reviewed here was a legitimate attempt at being an exception to this trend, but numerous and significant shortcomings in their study nullify it as a scientifically sound effort. Additionally, our assessment of Carter and Snow (2004) underscores the long-established problems associated with classical test theory and argues that modern methods of test construction and validation are needed in tandem with more sophisticated research designs in order to advance the field of relationship science and to increase confidence in research findings.

Aside from improved methodologies, we expect that if online testing is to become successful, online testing services – including online

matchmaking – will soon find themselves in the same situation as the publishers of such well-known tests as the GRE, MCAT, LSAT and GMAT. Embretson (1999) noted that although most textbooks continue to emphasize classical test theory, professional and commercial psychological measurement has rejected the classical approach in favor of methods derived from Item Response Theory (IRT) and variants thereof, such as Rasch (1960/1980) scaling (cf. Bond & Fox, 2001; Wright & Stone, 1979). In virtually all instances of large-scale testing where results have real-world implications for individuals completing the assessments, the use of classical test theory has been abandoned. The quality of computer dating services (or any online testing company for that matter) is only as good as the quality of their tests and methodologies. We expect therefore that it is only a matter of time before the advantages of IRT-based methods begin to be reflected in the practices of online testing services, just as they have been adopted in other testing domains.

In this context, it seems appropriate to introduce the differences between classical test theory and IRT using the four "rules" provided by Embretson (1995; 1999, p. 12):

1. The standard error of measurement differs between persons with different response patterns but generalizes across populations.
2. Shorter tests can be more reliable than longer tests.
3. Comparing test forms across multiple forms is optimal when test difficulty levels vary across persons.
4. Unbiased estimates of item properties may be obtained from unrepresentative samples.

In other words, the notion that all test scores are equally reliable has been abandoned in favor of local (i.e., level-specific) standard errors of estimate (*SE*). That is, no longer is there a single index of score reliability. Also, contrary to common wisdom, longer tests are not necessarily "better," as – depending on the variation in the trait levels of the respondents – many questions are almost guaranteed to be redundant. Rather, by using items that best address respondents' different trait levels (i.e., by purposely using *non*-parallel forms) greater measurement precision can be obtained. In the extreme, items are *selected* specifically to optimize reliability (minimize *SE*). When this is done in an interactive, computerized fashion one speaks of Computer Adaptive Testing, or CAT (Wainer, 2000). The savings achieved in the number of items needed when using CAT methods typically approaches 50%. Further, given an appropriate item-pool, *smaller SE*s can be achieved using CAT than with fixed-length tests. Even greater savings may obtain when the main

objective is to classify respondents into a small number of mutually exclusive categories (Eggen, 2004).

Rasch (1960/1980) scaling is especially useful in providing extensive indices of model fit for items as well as persons, while covering a wide variety of data types, including binary items, rating scales, Poisson counts, percentages, and paired comparisons (Linacre, 2004). While item and person fit is certainly important, misfit is not a sufficient reason for rejecting the Rasch model. Rather, it should be understood that Rasch scaling provides a measurement model that specifies the conditions under which observations constitute trait measures. Accordingly, misfit is a property of the data, rather than the model. As Bond and Fox (2001) explained, "the goal is to create abstractions that transcend the raw data, just as in the physical sciences, so that inferences can be made about constructs rather than mere descriptions about raw data" (p. 3). Researchers are then in a position to formulate initial theories, validate the consequences of theories on real data, refine theories in light of empirical data, and follow up with revised experimentation in a dialectic process that forms the essence of scientific discovery. Conversely, misfit can be exploited in advanced areas of research where theory is sufficiently powerful to predict particular deviations from the Rasch model (Bond, 1995; Lange, Greyson, & Houran, 2004; Lange, Jerabek et al., 2004).

Lack of model fit is most problematic if people receive systematically different scores based *not* on corresponding trait differences, but rather on a group specific interpretation or understanding of the questions being asked (for a recent discussion, see: Lange, Irwin et al., 2000). Such biases may reflect culture-related differences in expressing one's feelings, opinions, or symptoms (Lange, Thalbourne, Houran, & Lester, 2002), or they may be related to respondents' age or gender (Lange, Thalbourne, Houran, & Storm, 2000). And, of course, it is only to be expected that tests translated into other languages may not yield measures equivalent to those obtained in the base language (van de Vijver, & Poortinga, 1997). Again, IRT methods are helpful in cases where relatively few items are affected by the translation, as it may be possible to recalibrate just these items. Similar issues play a role when old tests are adapted for online use, as the meanings of words sometimes change over time (for example, consider how the vernacular versions of "cool" and "hot" have fluctuated over the last decades). Finally, differences in the method of administration (i.e., offline vs. online) may systematically affect respondents' reactions to the questions. It is mandatory, therefore, that web adaptations of paper-and-pencil instruments be recalibrated based on online administrations. Moreover, the exact same item format and layout should be used during pilot testing and operational use, and online tests

should not rely on norms that were established by paper-and-pencil methods (Naglieri et al., 2004).

Along with a desire to avoid legal issues related to unsubstantiated public claims of reliability, validity, and efficacy, the present context provides many good technical reasons for a switch from classical test theory to IRT, especially since online testing typically produces sufficiently large datasets to expand the approach into many new directions. For instance, the flexibility afforded by web-based item administration easily allows for the introduction and calibration of new questions or question types, *without* losing continuity and without the need to re-compute baselines or previously established cutoff scores. Finally, given the availability of a variety of person fit statistics it may be possible to identify outlying (e.g., malingering or unmotivated) respondents (Wright & Stone, 1979).

Conclusions

The online dating industry is clearly growing in importance as an industry, not only because it is becoming a popular and efficient way for busy singles to find love-interests, but because of the rich and valuable information that it provides for potentially reducing the rising divorce rate and other types of unsuccessful relationships. Therefore, it is crucial that online matching services purporting to use empirically validated matching systems actually do validate their systems and release their findings to the public. Doing so will allow the public to make more informed decisions about which services to register for, and facilitate the development of increasingly error-free, higher quality matching systems.

Applying IRT and Rasch scaling methodologies to the development and validation of online compatibility testing – and disclosing those findings for public and academic scrutiny *without divulging proprietary information* – has been proven feasible by the relationship-building company TRUE.com. They have posted on their website *The Technical Manual for the TRUE Compatibility Test*™ (TRUE & Jerabek, 2004; see: http://true.com/images/tctmanual.pdf?svw=bnavbar2) that details the reliability and validity data for their testing methods and which was independently audited by a tests and measurements expert. This effort seems to currently set the standard for addressing the disclosure concerns as raised by Houran (2004) and Finn and Banach (2000).

Although it remains to be seen whether other companies follow this standard, the prospect that millions of singles are making life-changing decisions based on compatibility tests that are not scientifically sound is a sobering one. Indeed, medical patients would not take a drug that has not been approved by the FDA (unless they are desperate) and likewise people looking for relationships should not so willingly trust online

psychological tests and matching systems that have not been independently proven to meet professional testing standards. Thus, it is through careful research designs utilizing such IRT approaches that evidence for the efficacy of online matchmaking methods will be convincing, as well as yield new and provocative insights that will inform current thinking on the building blocks for successful relationships of all kinds.

REFERENCES

Ahuvia, A. C., & Adelman, M. B. (1992). Formal intermediaries in the marriage market: a typology and review. *Journal of Marriage and Family*, *54*, 452-463.

American Educational Research Association, American Psychological Association, & National Council on Measurement (2002). *Standards for educational and psychological testing*. Washington, DC: Author.

Aronson, E., Wilson, T. D., & Akert, R. M. (1994). *Social psychology: the heart and the mind*. New York: Harper Collins College Publishers.

Baker, A. (2002). What makes an online relationship successful? Clues from couples who met in cyberspace. *CyberPsychology & Behavior*, *5*, 363-375.

Beach, S. R. H., Whitaker, D., Jones, D.J., & Tesser, A. (2001). When does performance feedback prompt complementarity in romantic relationships? *Personal Relationships, 8*, 231-248.

Berscheid, E., & Reis, H. T.(1998). Attraction and close relationships. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology*, *Vol. 2*. (4th ed.) (pp. 193-281). New York: McGraw-Hill.

Bond, T. G. (1995). Piaget and measurement II: empirical validation of the Piagetian model. *Archives de Psychologie*, *63*,155-185.

Bond, T. G., & Fox, C.M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Bor, R., Prior, N., & Miller, R. (1990). Complementarity in relationships of couples affected by HIV. *Counselling Psychology Quarterly, 3*, 217-220.

Bramlett, M. D. & Mosher, W. D. (2002). *Cohabitation, marriage, divorce, and remarriage in the United States* (Vital and health statistics series 23, Data from the National Survey of Family Growth; no. 22). Hyattsville, MD: National Center for Health Statistics Department of Health and Human Services.

Buckwater, J. G., Carter, S. R., Forgatch, G. T., Parsons, T. D., & Warren, N. C. (2004, May 11). *Method and system for identifying people who are likely to have a successful relationship* (submitted May 11, 2000). U. S. Patent 6,735,568. Washington, DC.

Carroll, L., Gilroy, P., Hoenigmann-Stovall, N., & Turner, J. A. (1998). Sexual identities and complementarity in lesbian and gay male relationships. *Journal of Gay & Lesbian Social Services*, *8*, 1-12.

Carter, S., & Snow, C. (2004). *Helping singles enter better marriages using predictive models of marital success*. Poster presented at the 16th Annual Convention of the American Psychological Society, Chicago, IL, May 27-30.

Caspi, A., & Herbener, E. S. (1990). Continuity and change: assortative marriage and the consistency of personality in adulthood. *Journal of Personality and Social Psychology*, *58*, 250-258.

Cattell, R. B., & Nesselroade, J. R. (1967). Likeness and completeness theories examined by sixteen personality factor measures on stably and unstably married couples. *Journal of Personality and Social Psychology*, *7*, 351-361.

Christensen, A., & Heavey, C. L. (1990). Gender and social structure in the demand/withdraw pattern of marital conflict. *Journal of Personality and Social Psychology*, *59*, 73-81.

Cobb, R. J., Davila, J., & Bradbury, T. N. (2001). Attachment security and marital satisfaction: the role of positive perceptions and social support. *Personality and Social Psychology Bulletin*, *27*, 1131-1143.

Condon, J. W., & Crano, W. D. (1988). Inferred evaluation and the relation between attitude similarity and interpersonal attraction. *Journal of Personality and Social Psychology*, *54*, 789-797.

Dryer, D. C., & Horowitz, L. M. (1997). When do opposites attract? interpersonal complementarity versus similarity. *Journal of Personality and Social Psychology*, *72*, 592-603.

Edmonds, V. H. (1967). Marital conventionalization: definition and measurement. *Journal of Marriage and the Family*, *29*, 681-688.

Egan, J. (2003). Love in the time of no time. *The New York Times Magazine*, 23 November, p. 66.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Enschede: Ipskamp.

Embretson, S. E. (1995). The new rules of measurement. *Psychological Assessment*, *8*, 341-349.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S.E. Embretson and S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know* (pp. 1-16). Mahwah, NJ: Lawrence Erlbaum.

Finn, J., & Banach, M. (2000). Victimization online: the down side of seeking human services for women on the Internet. *Cyberpsychology and Behavior*, *3*, 243-254.

Goot, D. (2004, March). Love, algorithmic style. *Wired Magazine*, *12.03*. http://www.wired.com/wired/archive/12.03/start.html?pg=6. Accessed April 20, 2004.

Gottman, J. M. (1994). *What predicts divorce? the relationship between marital processes and marital outcomes*. Hillsdale, NJ: Erlbaum.

Gottman, J. M., Murray, J. D., Tyson, R., & Swanson, K. R. (2002). *The mathematics of marriage: dynamic nonlinear models*. Cambridge, MA: MIT Press

Gross, R. & McIlveen, R. (1998). *Psychology: a new introduction*. London, Hodder & Stoughton.

Holman, T. B. (2001). *Premarital prediction of marital quality or break up: research, theory, and practice*. New York: Kluwer.

Houran, J. (2004). *Ethics in cross-cultural compatibility testing in Europe: an opportunity for industry growth*. Paper presented at the Internet Dating /

Online Social Networking Industry Association Inaugural Meeting, Nice, France, July 15-16, 2004.

Houran, J., & Lange, R. (2004). Redefining delusion based on studies of subjective paranormal ideation. *Psychological Reports*, *94*, 501-513.

Houts, R., Huston, T.L., & Robins, E. (1996). Compatibility and the development of premarital relationships. *Journal of Marriage and the Family*, *58*, 7-20.

Hunsley, J., Pinsent, C., Lefewbvre, M., James-Tanner, S., & Vito, D. (1995). Construct validity of the short forms of the Dyadic Adjustment Scale. *Family Relations*, *44*, 231-237.

Huston, T. L., Caughlin, J. P., Houts, R. M., Smith, S.E., & George, L. J. (2001). The connubial crucible: newlywed years as predictors of marital delight, distress, and divorce. *Journal of Personality and Social Psychology, 80*, 237–252.

Huston, T. L., Niehuis, S., & Smith, S. E. (2001). The early marital roots of conjugal distress and divorce. *Current Directions in Psychological Science, 10*, 116-119.

Ickes, W. (Ed.) (1985). *Compatible and incompatible relationships*. New York: Springer-Verlag.

Karney, B. R., & Bradbury, T. N. (1995). The longitudinal course of marital quality and stability: a review of theory, methods, and research. *Psychological Bulletin*, *118*, 3-34.

Kerckhoff, A. C. and Davis, K. E. (1962). Value consensus and need complementarity in mate selection. *American Sociological Review*, *27*, 295-303.

Klohnen, E. C., & Mendelsohn, G. (1998). Partner selection for personality characteristics: a couple-centered approach. *Personality and Social Psychology Bulletin*, *24*, 268-278.

Kurdek, L. A., & Schmitt, J. P. (1987). Partner homogamy in married, heterosexual cohabiting, gay, and lesbian couples. *Journal of Sex Research*, *23*, 212-232.

Lange, R., Greyson, B., & Houran, J. (2004). A Rasch scaling validation of a 'core' near-death experience. *British Journal of Psychology*, *95*, 161-177.

Lange, R., Jerabeck, I., & Houran, J. (2004). *Building blocks for satisfaction in long-term romantic relationships: evidence for the complementarity hypothesis for romantic compatibility*. Paper presented at the Adult Development Symposium Society for Research in Adult Development Preconference, AERA, San Diego, CA, August 11.

Lange, R., Irwin, H. J., & Houran, J. (2000). Top-down purification of Tobacyk's Revised Paranormal Belief Scale. *Personality and Individual Differences*, *29*, 131-156.

Lange, R., Thalbourne, M. A., Houran, J., & Lester, D. (2002). Depressive response sets due to gender and culture-based differential item functioning. *Personality and Individual Differences*, *33*, 937-954.

Lange, R., Thalbourne, M. A., Houran, J., & Storm, L. (2000). The Revised Transliminality Scale: reliability and validity data from a Rasch top-down purification procedure. *Consciousness and Cognition*, *9*, 591-617.

Levinger, G. (1986). Compatibility in relationships. *Social Science*, *71*, 173-177.

Linacre, J. M. (2004). *Facets Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Mand, A. (2004). Dr. Love is in. ABCNEWS.com. http://abcnews.go.com/sections/US/Relationships/online_dating_040326-1.html. Accessed April 20, 2004.

Mulrine, A. (2003). Love.com. *US News & World Report*, September 29, pp. 52-58.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: new problems, old issues. *American Psychologist*, *59*, 150-162.

Neff, L. A., & Karney, B. R. (2003). The dynamic structure of relationship perceptions: differential importance as a strategy of relationship maintenance. *Personality and Social Psychology Bulletin*, *29*, 1433-1446.

Nowicki, Jr., S., Yaughn, E. (1999). Close relationships and complementary interpersonal styles among men and women. *Journal of Social Psychology*, *139*, 473-478.

Pilkington, C. J., Tesser, A., & Stephens, D. (1991). Complementarity in romantic relationships: a self evaluation maintenance perspective. *Journal of Social and Personal Relationships*, *8*, 481 504.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: MESA Press.

Scheier, M. F., & Carver, C. S. (1992). Effects of optimism on psychological and physical well being: theoretical overview and empirical update. *Cognitive Therapy and Research*, *16*, 201-228.

Sharpley, C. F., & Cross, D. G. (1982). A psychometric evaluation of the Dyadic Adjustment Scale. *Journal of Marriage and the Family*, *7*, 739-741.

Spanier, G. B. (1976). Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family*, *38*, 15-28.

Spanier, G. B., & Cole, C. L. (1976). Toward clarification and investigation of marital adjusment. *International Journal of Sociology of the Family*, *6*, 121-146.

Spanier, G. B. & Thompson, L. (1982). A confirmatory analysis of the Dyadic Adjustment Scale. *Journal of Marriage and the Family*, *44*, 731-738.

Sternberg, R. J. (1986). A triangular theory of love. *Psychological Review*, *93*, 119-135.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, *55*, 293-326.

Sussman, S., & Reardon, K. K. (1987). Asset equality or similarity as determinants of perceived marital effectiveness: a rules perspective formulation. *Representative Research in Social Psychology*, *17*, 37-52.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, *103*, 193-210.

TRUE, LLC, & Jerabek, I. (2004). *The technical manual for the TRUE Compatibility Test (TCT)*™. Irving, TX: Author.

van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 29-37.

Wainer, H. (2000). *Computerized adaptive testing: a primer*. Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Whitty, M. T. (2003). Cyber-flirting: playing at love on the Internet. *Theory and Psychology*, *13*, 339-357.

Whitty, M. T., & Carr, A. N. (2003). Cyberspace as potential space: considering the web as a playground to cyber-flirt. *Human Relations*, *56*, 869-891.

---

[i] "Online Coaches Help Refine Personal Ads," *Asian Wall St. Journal*, at A8 (Aug. 17, 2004) (citing research conducted by Marketdata Enterprises, Inc.).

[ii] "Yahoo Adds Video, Voice to Online Dating Service," SilconValley.com (Jan. 12, 2003). http://www.siliconvalley.com/mld/siliconvalley/4933379.htm

[iii] "Online Coaches Help Refine Personal Ads," at A8 (citing research conducted by Nielsen//NetRatings).

[iv] PR Newswire, "Internet Dating: Boom Time for Clicking Online" (Feb. 9, 2004).
http://www.findarticles.com/p/articles/mi_m4PRN/is_2004_Feb_9/ai_113042296

[v] http://www.eharmony.com/core/eharmony?cmd=community-background

[vi] http://corp.match.com/news_center/nc_at_a_glance.aspx

[vii] http://www.perfectmatch.com/Images/tour/slide1.gif

[viii] http://www.perfectmatch.com/Images/tour/slide2.gif

[ix] Standard deviations were derived from Carter and Snow (2004) by multiplying the *SE* values listed in their Table 2 by the square root of the groups' respective sample sizes. The results were then squared to obtain the variances shown here.

Authors' notes: We thank David Reid for his assistance in the preparation of this paper and the reviewers for their helpful comments.