University of Nevada, Reno

Exploring the Predictive Utility of Implicit Relational Assessment Procedure (IRAP) with Respect to Performance in Organizations

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology

by

Gregory S. Smith

Dr. Ramona Houmanfar/Dissertation Advisor

December, 2013

University of Nevada, Reno
Statewide • Worldwide

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

GREGORY S. SMITH

entitled

Exploring The Predictive Utility Of Implicit Relational Assessment Procedure (IRAP) With Respect To Performance In Organizations

be accepted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Ramona Houmanfar, Ph.D., Advisor

Mark Alavosius, Ph.D., Committee Member

Steven Hayes, Ph.D., Committee Member

Barbara Kohlenberg, Ph.D., Committee Member

Sushil Louis, Ph.D., Graduate School Representative

Marsha H. Read, Ph. D., Dean, Graduate School

December, 2013

Abstract

The Implicit Relational Assessment Procedure (IRAP) has been used as a means of measuring implicit attitudes, or assessing implicit verbal relations, for several years. Much of the early work with IRAP consisted of validating the results obtained using the IRAP with those of another well-documented tool used for measuring implicit attitudes (i.e., IAT; see Barnes-Holmes, Barnes-Holmes, and colleagues). More recently, another iteration of the IRAP instrument, the Mixed Trial-IRAP (MT-IRAP), has emerged and early work has compared its results to those of previous studies employing the traditional IRAP instrument. Findings thus far have been promising, leading researchers to begin asking the next logical set of empirical questions; primarily, to what extent are the measures captured by the IRAP instruments indicative or predictive of more overt, probable patterns of behavior in naturalistic settings, such as the home, the workplace, or the community at large. Recent work has begun to address this question and more research is needed. The present study investigated this question as it related to patterns of behavior in organizational settings by asking participants to complete both of the two IRAP instruments comprising target stimuli related to the workplace and workplace behavior, in addition to exposing participants to an analog data entry work task, with dependent measures related to those concepts assessed in the IRAPs. IRAP results were then correlated with the overt behavior patterns from the analog work task to evaluate the extent to which the IRAP results were predictive of such behavior, in this particular (i.e., organizational) setting.

Table of Contents

List of Tables

List of Figures

Exploring the Predictive Utility of IRAP with Respect to Performance in Organizations

The concepts of implicit attitudes and biases are relatively new in the field of psychological science. Well-known theory and research pertaining to implicit attitudes and biases extends back only a couple of decades. The initial work in this line of research originates from the social psychological wing, with an emphasis on cognitively oriented psychological theories of human cognition as its foundation. The most prominent early work in this area was that of Greenwald and colleagues (e.g., Greenwald & Banaji, 1995; Greenwald, McGhee & Schwartz, 1998,), which utilized an instrument known as the Implicit Association Test (IAT) to measure implicit attitudes and biases, as defined by these researchers. The procedural details of IAT are informed by the cognitive psychological theory of associationism, which underlies the conceptualization of implicit attitudes (also known as implicit cognitions or implicit biases) and thus the means by which one could measure such attitudes. Following is a brief account of implicit biases and the IAT, based on social-cognitive psychological theory, which preceded the development of the Implicit Relational Association Procedure (IRAP), with which the present study is concerned.

Implicit Attitudes and the Implicit Association Test (IAT)

The seminal research publication in this line is Greenwald, McGhee, and Schwartz (1998), which outlines the authors' conceptual definition of implicit attitudes and the assessment tool developed to measure such attitudes (i.e., IAT). These researchers defined implicit attitudes as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects” (Greenwald & Banaji, 1995, p. 8). In accordance with this definition, it was suggested that implicit attitudes manifest as actions or judgments which are automatically activated through the activation of associations between mental representations of stimuli contacted in the environment and other stimuli or concepts stored in memory (Greenwald et al., 1998; Hughes, Barnes-Holmes & De Houwer, 2011). Based on this perspective, associations are considered to be links between mental representations of various stimuli and concepts, which have been encountered in the environment and presumably paired in some spatiotemporal manner, and which have been encoded and stored in memory as such. As a result, when one stimulus is encountered at a later time in the environment, it acts as a trigger in activating the mental representation of the second stimulus with which it is associated (Greenwald et al., 1998; Hughes et al., 2011; Shanks, 2007). In order to attempt to measure the extent to which different concepts are associated in memory, Greenwald and colleagues developed the IAT.

The IAT is a computer-based task which measures response latencies and involves mapping or pairing certain concepts together (e.g., Flower→Positive) on one response key and two different concepts (e.g., Spider→Negative) on the other response key. Response latencies are measured for each response in which a flower word and positive word are associated within the task and for each response in which an insect word and negative word are associated. Following that task, a second task is completed in which the paired concepts are reversed, for example, a flower word and a negative word mapped to one response key and an insect word and a positive word mapped to the other key. Response latencies are measured for each of these associated responses and compared to the response latencies from the first task. The result tends to be that individuals respond more quickly when flower words are paired with positive words and when insect words are paired with negative words, relative to when the pairings are reversed (i.e., flower and negative words, and insect and positive words; Greenwald et al., 1998).

Researchers in the social-cognitive domain operating from the perspective of associationism take the behavioral results described above and make the assumption that these IAT effects are a proxy or index of the extent to which these stimuli are linked and associated in memory as mental representations. In addition to this assumption, it is also assumed that the mental associations of stimuli in memory are the cause of the behavioral responses (i.e., shorter or longer response latencies) observed in IAT performances (Hughes et al., 2011).

Although the associationistic account described above has been the dominant perspective adopted within the area of implicit attitude research, it rests entirely on hypothetical constructs, such as internal storage of memory, associations or links among stored memories, and other mental processes, which are in no way observable or testable and therefore not amenable to direct scientific support or refutation. Instead, researchers are only able to observe the environmental context, such as how the IAT task is arranged, and the behavior of the participant, measured in terms of response latencies. The remainder of the associationistic account of implicit attitudes, their activation, and their effect on observable behavior is merely inferred from the observable behavior and amounts to nothing more than hypothetical intervening variables.

The foregoing discussion of implicit attitudes and human language and cognition, from a cognitive-associationistic perspective, is one of several ways such complex human phenomena as language and cognition have been conceptualized and investigated. Those of a behavioral rather than cognitive psychological persuasion have developed a contemporary account of human language and cognition based on behavioral theory and principles. As may be expected, this theory, known as Relational Frame Theory (RFT, see Hayes, Barnes-Holmes & Roche, 2001, for a detailed discussion), differs markedly from the cognitive-associationistic theory presented above, as does the conceptualization of implicit attitudes and how one could proceed with measuring and investigating them, accordingly. Importantly, RFT does not appeal to hypothetical constructs and mediating processes and is instead based on principles of behavior science, which have been abstracted from repeated scientific observations (i.e., things and events which are observable and available for confirmation) of behavior-environment relations. It is to a discussion of this behavioral theory and conceptualization of implicit cognitions that we now turn.

Relational Frame Theory (RFT)

Rather than basing the foundation of complex human behavior on associationistic forms of learning and responding, RFT instead bases complex human behavior and learning on relational responding, as operant behavior. In its simplest form, relational responding simply refers to responding to one stimulus in terms of another stimulus (Hayes et al., 2001). As an elementary example, an organism can be taught, based on principles of operant learning, to choose the larger of two circular objects. In one particular instance, one circle has a diameter of 18 mm, while another circle has a diameter of 21 mm. In this instance, an organism that has been exposed to an appropriate learning history will correctly pick the circle with the 21-mm diameter. However, in another particular instance, in which the circle with the 21-mm diameter is presented in the context of a circle with a 24-mm diameter, that same organism will correctly choose the circle with the 24-mm diameter, as opposed to the 21-mm diameter circle, which was the correct choice in the previous instance.

This sort of relational responding, in which the critical properties of the relation among the various stimuli are based on the formal, physical properties of the objects (e.g., physical size or diameter, in the above example), has been demonstrated with a number of organismic species and is not unique to human behavior (e.g., Harmon, Strong & Pasnak, 1982). What is unique to human behavior, however, and is held to be at the core of complex human behavior such as language, reasoning, and logic, is what is known as *arbitrarily applicable relational responding* (AARR; Hayes et al., 2001). As suggested by the name, AARR refers to the capability of humans to relationally respond to various stimuli based on properties attributed to the stimuli in accordance with acquired cultural convention (i.e., arbitrarily), rather than properties having to do with the physical form of the stimuli. Arbitrarily applicable relational responding appears to be a uniquely human psychological phenomenon which underlies much of the conventional forms of responding observed in human cultures.

As an example of AARR, consider again the three circles of differing diameters from the previous example. However, in the present example, consider the implications of the additional information that the 18-mm diameter circle is actually a dime, the 21-mm diameter circle is a nickel, and the 24-mm diameter circle is a quarter. Assuming the appropriate learning history to which members of the U.S.A. are typically exposed, if an adolescent or adult was presented with the 18-mm diameter coin and the 21-mm diameter coin and asked to choose the one that is bigger, we may see a number of different responses. Foremost, the individual in question may ask for further clarification, in terms of whether the question is which is “bigger” or which is “worth more.” If we specify that we are interested in which is bigger, in terms of physical size, then the person will correctly choose the nickel, as in the previous example of relational responding. However, if we specify that the subject is to choose the one that is “greater than” or “more,” in terms of value, then the subject will correctly choose the dime.

The concept of value, as imbued and trained by the culture—in this case the culture of the U.S.A.—and the correct choice of the dime in the context of the physically larger nickel, is an example of arbitrarily applying relational responding in this particular context, in which the arbitrarily applied relation is that of “greater than” in terms of “value.” The function of greater value has nothing to do with the physical or formal properties of which is larger and is instead based upon cultural convention, in which the dime is said to be of greater value and to which members of the socio-cultural community adhere, thus arbitrarily ascribing greater value to the dime than the nickel.

It is clear, then, that arbitrarily applied relations must be appropriately shaped and taught by the social community and members of the community must learn to respond appropriately to such conventions. Evidence of this can be seen when young children, who have not appropriately learned the concept of value and how it is applied to coinage, will happily choose nickels over dimes, based on the formal property of the size of the nickel relative to the dime. The most obvious example of AARR with respect to various stimuli comes in the instance of language. This is so, given that words are stimuli that are not solid physical objects per se, in the sense that a coin or a rock or a human are solid physical objects. Words are stimuli that exist in our human verbal behavior and only have meaning insofar as it is imbued and ascribed by a particular socio-cultural group of humans, and the extent to which that particular group of humans conventionally responds to those words. The formal properties of words as stimuli refer to the letters and sounds that constitute the words, as they are either vocally uttered through the aural medium or written through the visual medium. The formal properties of words are in no way necessarily tied to the meaning of the words, other than at some point in time people of a given socio-cultural group ascribed a particular meaning to a particular combination of sounds and ultimately symbols (e.g., letters of an alphabet), which has been maintained by the lineages of that group of persons and persisted through their culturally conventional responding. As a result, arbitrarily applicable relational responding can be applied to physical objects through human referential language with respect to those objects (as in the case of the value of dimes and nickels), as well as to words and concepts that exist within our language.

At this point the interested reader may ask, “How does one actually learn the ‘meaning’ of words as stimuli in the first place?” Relational Frame Theory posits that this kind of learning occurs in the same manner as do other forms of operant learning observed with both human and nonhuman organisms. In the simplest, earliest stages of learning the arbitrarily applied meanings of words, learning proceeds in the form of naming, in which words that refer to concrete, physical objects, such as coins and rocks, are taught using basic principles of operant learning (Hayes et al., 2001). For example, a mother teaching her young child will engage in teaching naming skills with a variety of stimuli, including the child’s toys. The mother may hold up the child’s ball and say the word “ball.” This also occurs with a myriad of other objects in the home, including “water,” “potty,” and even the parents themselves, “mama” and “dada,” which often may be the first successfully mastered instances of naming on the part of the child. Although this process may begin slowly and the child may not correctly name stimulus objects quickly, as the child begins to emit appropriate vocal utterances they are met with praise and visible excitement from the parent, which is assumed to be reinforcing for the child. As the child begins to emit more and more vocal responses that may (or in some cases may not at all) resemble the appropriate form of the word in question, the child may receive more praise from the parents and the parents will continue to teach more and more names.

Understanding the process of naming by young humans has its roots in behavior analytic research into a behavioral phenomenon known as stimulus equivalence (Sidman, 1986). In the research tradition of stimulus equivalence it has been demonstrated that, given an appropriate history of training, an organism will come to respond to one stimulus as though it is “equivalent to” another stimulus. In early equivalence research, this was demonstrated utilizing the research paradigm known as matching-to-sample, a preparation in which a single “sample” stimulus is presented along with an array (typically three) of other “comparison” stimuli (Sidman & Tailby, 1982). In this instance, in the presence of a given sample stimulus, A, the organism should choose only one of the three available comparison stimuli, B1, B2, or B3. If the correct choice (e.g., B1) is made, some form of reinforcement occurs and on subsequent occasions when that particular sample stimulus A is again presented, the organism will emit the appropriate choice response, B1.

Such training has been successfully completed with many different species of organisms. Of particular interest, however, is how some species respond when the situation is modified in a particular way. Specifically, if the comparison stimulus, B1, which previously served as the correct choice response among the array of comparison stimuli, is now presented alone as the sample stimulus and the stimulus A, which originally served as the sample stimulus, is now presented in an array of different comparison stimuli (e.g., as A1, in addition to A2 and A3), some organisms will choose the original sample stimulus A1 out of the current array of comparison stimuli without ever having had any exposure to or training with this particular set of conditions. Although this may seem to humans to be fairly obvious and not particularly surprising, as it turns out humans are some of the only organisms that naturally and consistently engage in such behavior. This pattern of responding, in which reversing the contingency of reinforcement, in terms of antecedent stimuli and appropriate choice response without the appropriate response having been previously trained, is known in RFT as *mutual entailment* and is an example of *derived relational responding* (DRR; Hayes et al. 2001). In this particular case, the relational response, which we may describe as “in the presence of A1, choose B1,” is reversed and subsequently takes the form of “in the presence of B1, choose A1,” although the latter circumstance has never been previously encountered or trained. Essentially, the two stimuli have become equivalent, in terms of how the organism responds to them, and part of that equivalence relation, specifically, the “in the presence of B1, choose A1" relation, has been derived on the part of the organism, as it has not been explicitly trained.

Derived relational responding is also observed in other forms of relational responding described by RFT, for example, in responding referred to as *combinatorial entailment* (also referred to as *combinatorial mutual entailment*; Hayes et al., 2001). Building on the previous example of stimuli A1 and B1, we can introduce another stimulus, C1. We can proceed with the training exactly as it is described above, in which relational responses we may describe as, "in the presence of A1, choose B1," and, "in the presence of A1, choose C1," are directly and explicitly trained. As previously discussed, without any direct training, many organisms will derive the reversed relation (i.e., mutual entailment) and correctly choose A1 in the presence of B1 or A1 in the presence of C1. However, typically only humans will derive another additional relation; specifically, without having had any prior exposure to such conditions, humans will choose C1 (among C1, C2, and C3) in the presence of B1 or vice versa (i.e., choose B1 in the presence of C1). Again, this behavioral phenomenon is referred to as combinatorial entailment and in elementary and middle schools across the planet commonly takes the form of: if $x = y$ and $x = z$, then $y = z$ (and $z = y$; also referred to as the *law of transitivity* in mathematics).

Lastly, yet another form of DRR as it is observed in relational responding is articulated in RFT and it is known as *transformation of stimulus function* (Hayes et al., 2001). Specifically, transformation of stimulus function refers to the behavioral phenomenon whereby a stimulus acquires the function of another stimulus with which some kind of relational responding has been established. As an example of transformation of stimulus function, consider a situation in which two individuals are involved in a romantic relationship. A part of their behavioral repertoire as members of the relationship is to frequent the restaurant at which they had their first date. As such, the restaurant in particular and their visits to it carry a strong, positive or appetitive function; that is to say, they both enjoy going to the restaurant together, find it pleasing and appealing, and continue to do so for a long time, by which we may assume it is a reinforcing event for each of them. However, their relationship eventually comes to an end when one partner discovers that the other has been carrying on an affair with another person. In all likelihood, the partner who has been cheated on and left heartbroken will no longer visit the restaurant where they had their first date. This is due to the fact that the former partner has taken on new psychological functions as a result of the unfortunate circumstances of the breakup, and the restaurant has acquired a similar psychological function as that which the former partner has.

Said in more technical language consistent with RFT, the former partner and the restaurant were in many ways participating in relations of equivalence, which in RFT terms is more precisely known as a relation of coordination (i.e., similarity). As such, the restaurant had acquired stimulus functions that were in many ways similar to or coordinate with the psychological stimulus functions of the partner; namely, pleasant, desirable, appealing, etc. However, following the experience of the circumstances of the break up, the former partner acquired many new stimulus functions for the heartbroken individual, perhaps such as anger, displeasure, hurt, shame, and avoidance. Transformation of stimulus function is seen in this instance when the restaurant, although itself not *directly* involved in any particular experiences that would lead to its acquisition of unpleasant, avoidant stimulus functions, nevertheless acquires similar stimulus functions as those of the former partner (hence, *derived* relational responding). As a result of the relation of coordination having been established between the partner and the restaurant, the restaurant now becomes a stimulus which evokes the same behavioral responses of displeasure, shame, and avoidance through the transformation of stimulus function of other stimuli (i.e., the partner) that participate in a relational network with it (i.e., other stimuli with which there is some established history of relational responding).

It is worthwhile, at this point, to discuss the concept of stimulus function. Importantly, stimulus function can be distinguished from stimulus object. As a brief example, a stimulus object is a bicycle. The object itself has a particular spatiotemporal size, shape, and set of formal, physical properties. Though the physical properties of the bicycle may not change over time, the psychological stimulus functions of that stimulus object can. Consider that a young child who is enjoying his first bicycle rides it around constantly. At this point in the example, the bicycle probably has acquired many appetitive, pleasant stimulus functions, including serving as a source of enjoyment, a means of transportation to go see friends and to go other places, and perhaps even a source of admiration from friends, if it is a "cool" bicycle. Upon most occasions, the child probably cannot wait for an opportunity to go ride his bicycle. However, on one particular day, the child suffers a serious and painful injury as a result of riding the bicycle. As a result of this experience, the bicycle most likely acquires some new stimulus functions. At the very least, the child now approaches riding the bicycle with some caution, perhaps even fear. In a serious example, the child may no longer want to ride the bicycle at all or even want to own the bicycle (i.e., give it away or have his parents dispose of it). This is not to say that the previous stimulus functions of the bicycle have been “erased” or have disappeared, but for the moment the additional stimulus functions which have been acquired by the bicycle are those functions which presently influence the child’s behavior.

It should be made clear that the stimulus object, in this case a bicycle, does not change in its physical properties and the variety of stimulus functions that are built up through interactions with the bicycle are not stored in the bicycle itself (which probably seems obvious). More importantly, however, is recognition that the stimulus functions are neither stored in the child. Instead, the stimulus functions said to have been “built up” through experience exist only in the child’s psychological interactions with the bicycle. As such, stimulus and response functions are to be conceptualized as a singular unit, consisting of both the stimulating of the object and the responding of the psychological organism (Kantor, 1958), in this case the bicycle and child, respectively.

It should be clear in the foregoing example that if another child with no prior history of traumatic interactions with that or any other bicycle were to see it being given away (due to the first child no longer wanting it), the bicycle would have no such fearful stimulus functions and there would be no corresponding response functions on the part of the second child that would result in the second child avoiding and being afraid of the bicycle. Instead, that child would probably happily take the bicycle to be his own—especially if it were a “cool” bike. Interestingly, however, is the scenario in which the second child inquires of the first child why the bicycle is being given away. Consider the circumstance in which the first child recounts the horrific experience of having been thrown from the bicycle and breaking a limb or losing teeth or what have you (imagine your own gruesome version of the incident), and in vivid detail describes the pain and suffering of the experience.

It is not incredible to think that the second child may suddenly be less enthusiastic about taking that bicycle home and keeping it for himself. Although that second child has had no personal experience as yet with this bicycle, through the processes of verbal behavior, including mutual and combinatorial entailment with arbitrarily applicable and derived relational responding, along with transformation of stimulus function, that bicycle may acquire for the second child some of the stimulus functions it has for the first child. This example illustrates how any number of stimuli may enter into a vast number of relations with respect to other objects and concepts (i.e., stimuli as physical objects and words and concepts, each of which may acquire numerous different functions over time), without the psychological organism having had any direct experience with the objects and concepts being referenced. Again, it should be noted that the relations do not exist anywhere other than in the psychological interactions of the organism with a stimulus and are nothing more than descriptions of those interactions.

The foregoing example demonstrates how a single stimulus object (e.g., a bicycle) can come to acquire a number of different stimulus functions based on multiple experiences with the stimulus object, including direct contact with the object as well as contacting the object indirectly through verbal descriptions and rules which reference the object. RFT describes how various words as stimuli in our language used by a speaker specify the appropriate relational responses to be emitted on the part of a listener, when verbal descriptions and rules are utilized. In order to explain these technical concepts, the example of the nickel and dime will be appealed to once more.

Once members of the culture have learned the appropriate "meanings" of certain words and concepts as they exist in the language of the culture, there are words as stimuli which serve as contextual cues and serve to occasion certain instances of AARR on the part of the listener, in accordance with the current context. For example, when asked, "Which one is worth more?" in the case of the nickel and dime, the word "more" serves as a *relational contextual cue* (C_{rel}), which specifies that the relation of "more than," "larger than," or "greater than" (all of which may be used interchangeably, assuming an appropriate learning history) should be applied to the stimuli in question (Hayes et al., 2001). In just the presence of "more than" or "greater than," however, it may still not be clear which is the appropriate response on the part of the listener, as it is not clear whether the individual should respond to the two stimuli in terms of physical size or some other (possibly arbitrary) property.

In the case of the above example, when someone asks "Which is worth more?" the term "worth" serves as a *functional contextual cue* (C_{func}), which specifies the functional properties of the stimuli (i.e., one stimulus compared to another in terms of what), to be applied in the relational response (Hayes et al., 2001). The term "worth," given an appropriate learning history, is functionally equivalent to the term "value," which is the appropriate property of the stimuli in question to which the listener should respond. To the extent that the listener in this example has not learned the subtle relations among terms such as "worth" and "value," a more explicit and precise framing of the question would resemble, "Which coin has a value that is greater than the other?" By contrast, the question, "Which is larger in size?" still specifies the same relation (i.e.,

C_{rel} ; larger than or greater than), however, it incorporates a different C_{func}, which instead specifies the function of formal size of the stimuli.

Implicit Relational Assessment Procedure (IRAP)

As indicated in the name, the IRAP is an assessment tool that assesses relational responding on the part of a human participant. Additionally, the IRAP assesses relational responses which are referred to as *implicit* responses. Implicit responses in this case can be differentiated from explicit responses in terms of the measurement of such forms of responding (Barnes-Holmes, Barnes-Holmes, Power, Hayden, Milne & Stewart, 2006); Hughes et al., 2011). Explicit responding is conceptualized as when an individual is asked to provide a response to some set of stimuli, perhaps a question or survey, and the individual provides the response with sufficient time to consider the response. Explicit responding, in terms of its role in assessing attitudes on the parts of individuals, typically takes the form of answering surveys or questionnaires through vocal or written responses. In these instances, questions are posed to an individual and the individual has time to consider the answer to be provided, as is typically the case in most settings.

The limitation with explicit forms of responding in these cases is that individuals may not always provide "accurate" answers and instead may provide responses that appear to the individual to be most appropriate, or socially desirable, given the current context (Hughes et al., 2011). For example, someone applying for a sales job may be asked if he "likes people," to which the applicant will presumably reply that he does, since the applicant is likely to derive the rule that answering in the negative may adversely affect his likelihood of landing the job. In this instance, the answer regarding liking people may or may not correspond with how this individual would explicitly state his attitude towards liking people under different circumstances, such as when confiding in a close friend.

Explicit measures are contrasted with implicit measures, in that implicit measures are by definition considered to be somewhat automatic and emitted rapidly, as opposed to following temporally extended consideration (i.e., extended relational responding in the context of additional contextual cues, such as a potential employer). RFT researchers in the behavioral tradition have described implicit responses as those which are brief, immediate relational responses (BIRRs; Barnes-Holmes, Barnes-Holmes, Stewart & Boles, 2010). The underlying theory of BIRRs suggests that when individuals are presented with particular stimuli they engage in brief, immediate relational responses with respect to those stimuli, and the relational response in which they engage is a function of the individual's prior history of learning and history of responding with respect to those particular stimuli, and is under contextual control of the environment in which the relational response occurs.

In the case of the IRAP, the stimuli in question among which the participant makes a brief, immediate relational response are words presented on a computer screen (see Figure 1). Typically, a given word, known as the categorical or sample stimulus, is presented at the top of the computer screen. This term is typically an evaluative term, such as Good, Bad, Pleasant, Unpleasant, etc. In a typical IRAP procedure, there are two categorical terms which appear at the top of the computer screen and quasi-randomly alternate across trials. In addition to the categorical stimuli are a number of stimuli which may represent certain concepts, specifically those concepts toward which implicit attitudes are being measured, known as target stimuli. In some of the earliest IRAP work, which mapped onto early IAT work, the concepts involved were those of generally pleasant and unpleasant words or concepts. Specific target stimuli pertaining to the Pleasant concept class, for example, were Love, Caress, Freedom, Health, etc., and for the Unpleasant class were Abuse, Crash, Filth, Murder, etc. (Barnes-Holmes et al., 2006). Lastly, in the typical IRAP procedure, are the two response options available to the participant, and these represent the relational responses that must be made by the participant.

Essentially, given a particular target stimulus (e.g., Love) and a given evaluative stimulus (e.g., Pleasant), the participant must emit a relational response, for example "Similar" or "Opposite", in the context of the two stimuli (e.g., Love is—that is to say, similar to—Pleasant, in the case of the "Similar" response in the preceding example). The IRAP task can be conceptualized via RFT as presenting the participant with a C_{func} stimulus (i.e., the evaluative, categorical stimulus, such as "Good" or "Bad") along with the target concept of interest (i.e., the target stimuli). The participant must then emit a relational response (e.g., "Similar" or "Opposite"), which is a C_{rel} and specifies a relation among the evaluative stimulus (C_{func}) and target stimulus. The relational response is emitted by the participant by pressing one of two keys on the computer keyboard. Typically, the D and K keys are utilized and the particular relational response (e.g., Similar or Opposite) that corresponds to each key is indicated in the bottom corners of the computer screen (see Figure 1). The two available relational response options typically alternate quasi-randomly from one position (i.e., keyboard key) to the other over successive trials. Thus, the participant is required to make a relational response in terms of the target stimulus in the context of a particular evaluative, categorical stimulus and, most importantly, this response must be emitted very quickly. If the response is not emitted within a given latency (e.g., 2000 or 3000 milliseconds), a message to “Go faster!” appears on the screen. It is this particular aspect of rapid responding in the IRAP, as with the IAT, that is considered to drive the implicit nature of the responses being measured, as opposed to carefully considered, temporally extended responses, as is seen with explicit responding (Barnes-Holmes et al., 2006).

The extent to which key-presses on the part of a participant are reflective of BIRRs (i.e., implicit attitudes) is described through the Relational Elaboration Coherence (REC) model (Barnes-Holmes et al., 2010), which asserts that upon making visual contact with a given combination of categorical and target stimuli on a given IRAP trial, the participant engages in a covert BIRR, the probability of which is based largely on the participant’s learning history and current environmental context, and this BIRR may or may not correspond more or less closely to the overt key-press response which must be emitted under time pressure (Barnes-Holmes et al., 2010). The REC model contends that if the BIRR which is most likely to occur in the presence of a given combination of stimuli is more consistent with the relational response specified by a certain key-press response, then the key-press response should, on average, be emitted more quickly (i.e., shorter response latency). Conversely, if the BIRR which occurs in the presence of certain stimuli is less or not at all consistent with the relational response specified by the required key-press, then the key-press response should, on average, be emitted less quickly (i.e., longer response latency; Barnes-Holmes et al., 2010).

Hence, given a specific combination of categorical and target stimuli, the participant must, on different occasions, emit each of the relational responses available through the key press, such that an analysis of which relational key-press response tends to be emitted more quickly can be conducted. In order to accomplish this, the IRAP requires that in some instances the participant respond in one particular pattern (e.g., Pleasant→Love→Similar) and then in the opposite pattern in other instances (e.g., Pleasant→Love→Opposite). In the present example, if a participant were, over repeated presentations, to emit the "Similar" key press more rapidly (i.e., shorter response latency) relative to the "Opposite" key press, then it would be concluded that relationally responding to the concept of Love in terms of being similar to Pleasant is reflected in the history of learning and responding of the participant to a greater degree than is relationally responding to the concept of Love in terms of being similar to Unpleasant. In this way, it is said that the participant has an "implicit attitude" or "implicit bias" in terms of finding love to be pleasant.

The traditional IRAP preparation has been to develop two lists of target stimuli which can be said to loosely represent concept classes, to which implicit attitudes or biases are then measured using the instrument. Consider the following two examples, as noted previously, of broad concept classes incorporated in an IRAP (Barnes-Holmes et al., 2006). Pleasant: Love, Peace, Caress, Freedom, Health, Cheer; Unpleasant: Abuse, Sickness, Crash, Filth, Murder, Accident. In typical IRAP studies, the analyses of implicit attitudes toward target concepts does not occur at the level of the individual stimulus (e.g., Rose or Love), but rather at the level of the list of stimuli which are assumed to be functionally related (e.g., Love, Peace, Freedom, etc.). Additionally, the analyses tend to proceed not only at the level of the list of conceptually related stimuli, but also across groups of participants, such that every response latency for the relational response of Pleasant→[Love, Peace, Caress, Freedom, Health, Cheer]→Similar is averaged across all participants and compared to the average latency of responses for Pleasant→[Love, Peace, Caress, Freedom, Health, Cheer]→Opposite for all participants, in order to determine any pre-existing implicit attitude among the entire group of participants.

While the procedures and analyses described above reflect typical IRAP research and may have been useful early on for the purpose of validating the IRAP against the more well-established IAT, more recently there has been discussion of whether such an approach is best suited for measuring implicit attitudes on the parts of various individuals, as well as for any further purposes (e.g., applied uses) for which IRAP results might be utilized. Specifically, Levin, Hayes, and Waltz (2010) discuss several of these issues at length and also introduce a new iteration of the IRAP instrument, known as the Mixed Trial-Implicit Relational Assessment Procedure (MT-IRAP).

Mixed Trial-Implicit Relational Assessment Procedure

Levin et al. (2010) note that the traditional IRAP relies on a process of the experimenter choosing two lists of target stimuli words which are assumed to be semantically (and therefore functionally) related to the overarching concept classes being assessed in the procedure. However, doing so ignores the fact that each assessee comes to the assessment with a unique history of learning and responding to the particular stimuli in question. As an example, “Love” is often used to represent the concept class of “Pleasant,” however, it is quite possible that a given individual would not respond relationally to Love and Pleasant as being similar, based on that individual’s unique history of experiences. Consider the person who has lost the “love of his life,” who has experienced the “one who got away,” or the individual who has never been in love, has never known anyone to love him, and has come to view love as something that he will never be able to experience. In such a scenario, the word Love may come to acquire a very different stimulus function than for other individuals. Given such possibilities, it appears unwise to make such assumptions of similar functions of different stimuli across numerous individuals (Levin et al., 2010).

In the same vein, while it may be the case that a list of stimuli appear to be semantically related and may even be loosely functionally related for a given individual, when the response latencies for all of the stimuli within a list of a given concept class are averaged, any particular outliers, in terms of very strong attitudes or biases (as measured by largely differing response latencies across the two relational responses), are masked and therefore one cannot determine if there are any important differences among implicit attitudes toward the individual target stimuli within a given list. Depending upon the premise for use of the instrument and the extent to which the results inform other purposes, it may be critically important to be able to identify particularly strong, weak, or even opposing biases of a single stimulus relative to others in its presumed functional (i.e., list-level) class (Levin et al., 2010). Appropriately, then, Levin et al. argue for the importance of an IRAP analysis to be able to drill-down to the level of the individual stimulus, as well as the individual participant, such that the results may be used on an individual basis to the extent necessary, which would almost certainly be the case in applied work, if the IRAP were to be utilized in such a manner.

In order to accomplish the more robust individual-level analyses described above, the MT-IRAP incorporates a different procedural aspect, from which it derives the name

"Mixed Trial." Specifically, rather than forcing participants to respond in two opposing patterns on differing occasions (e.g., Similar on one occasion and Opposite on another occasion, in the presence of Pleasant→Love), the MT-IRAP allows participants to respond as they see fit, by adding another stimulus to the computer screen (see Figure 2). Specifically, the label "Truth" or "Lie" appears beneath the target stimulus and instructs participants to relationally respond to the stimuli in question, using the keyboard keys, based on their own explicit attitudes regarding the stimuli in question. In essence, the "Truth" and "Lie" labels serve as additional C_{func} cues which specify how a participant should relationally respond on given trial, based on the participant's explicit attitudes toward the stimuli in question. When the opposing Truth/Lie label appears on a subsequent trial involving the same stimuli, the participant should then emit the opposing relational response, as specified by the opposing C_{func} cue (i.e., Truth followed by Lie, or vice versa).

Building on the previous example used, a participant may encounter the combination of Pleasant→Love→Truth and have the relational response options of Similar and Opposite. For the participant who explicitly views Love as Pleasant, in the context of the Truth cue this participant will respond with the "Similar" key. In addition, when Pleasant→Love→Lie is presented, this same participant should then respond with the "Opposite" key. However, a different participant with a uniquely different history of experiences may instead choose to respond with the "Opposite" key when presented with Pleasant→Love→Truth. In this way, each participant will still emit the two different relational responses (e.g., Similar and Opposite), such that the appropriate analyses of response latencies can be conducted; however, it is up to the participant to judge on an individual basis how each target stimulus presented relates to the evaluative term presented and no *a priori* assumptions on the part of the assessor are made (Levin et al., 2010).

It should be noted that although the MT-IRAP incorporates the contextual cues of Truth and Lie, the experimenter should never assume that responses occurring under the contextual control of these cues in fact represent a participant's "True" feelings or not. In the case of socially sensitive topics, for example, racial or other prejudice, a participant who does not feel comfortable displaying socially unacceptable attitudes may still provide the socially acceptable response in the presence of the Truth label, even if the response does not reflect the participant's "true" explicit attitudes on the matter (e.g., what this participant would disclose to a close, trusted friend). Instead, the Truth and Lie labels serve to guide the participant's responding such that each relational response option will, at different times (i.e., in the presence of Truth or Lie), be emitted, in the same way the two different response patterns are forced in the traditional IRAP through explicit feedback and correction procedures.

While the conceptual grounds for the modifications encompassed within the MT-IRAP seem warranted based on the theory of RFT, there has yet to be any direct comparison of the two iterations of the instrument and, furthermore, there has been only one publication involving the MT-IRAP to date (Levin et al., 2010). And while both instruments are based on the same theory of human language and cognition, it is not yet clear whether one instrument yields more accurate, precise, or reliable results, or if the results of the two instruments are even convergent, for that matter. It is clear that more research incorporating the MT-IRAP is warranted and, further, it would seem that some research comparing the two instruments is also needed.

In addition to the question of whether either IRAP instrument is able to accurately and reliably measure implicit attitudes, as defined, a larger question looms in the literature surrounding all instruments said to measure implicit biases and the biases they are said to measure: namely, do implicit attitudes or biases in any way predict how an individual is likely to behave in naturalistic settings outside of the laboratory, where the biases being measured are more relevant? For example, if the individual applying for a sales position were to demonstrate an implicit bias in terms of not preferring to interact with others, does the implicit bias bear any prediction of or relation to how the individual is likely to interact with customers in a sales position? If the answer is yes, then hiring the individual may be cautioned against; however, if the answer is no, then there should be no cause for concern in hiring the individual for the position. It has been suggested in the literature on implicit bias that this question—whether implicit biases have any effect on or at least correlate with more persistent, overt patterns of behavior in naturalistic settings—is the main impetus behind the conceptualization of and research into the construct of implicit bias (De Houwer, 2002).

With this in mind, it is paramount that studies utilizing the IRAP begin to evaluate the extent to which implicit biases as measured by the IRAP inform, predict, correlate with, or influence other behavioral patterns of interest. Attempts to validate the IRAP using the “known-groups” approach, which incorporates different groups of participants (e.g., meat eaters and vegetarians) who are “known” (assumed) to exhibit specific attitudes (e.g., vegetarians find pictures of prepared meat dishes to be unpleasant), have been conducted and have generally found that implicit attitudes as measured by the IRAP correspond with the attitudes predicted of each known group (Barnes-Holmes, Murtagh, Barnes-Holmes & Stewart, 2010; Barnes-Holmes, Waldron, Barnes-Holmes & Stewart, 2009). The known-groups studies, however, are unable to directly answer the question of the extent to which an IRAP assessment can predict how an individual will behave under certain conditions and subsequently support those predictions with behavioral observations of the specific behaviors of interest. To date, there appear to be few studies that have made an explicit effort to begin to address this question. Specifically, Nicholson and Barnes-Holmes (2012) utilized the traditional IRAP to first investigate whether the IRAP was capable of differentiating among individuals with high or low spider fear and, secondly, to evaluate whether implicit spider fear, as measured by the IRAP, was predictive of approach behavior toward a live tarantula in the laboratory. Their results are promising in that D-IRAP scores pertaining to the Pro- and Anti-Spider trial types within the IRAP tended to correlate with and predict approach behavior in approximately 70% of cases in the study. The study takes an important first step in the direction of answering perhaps the most important question related to the IRAP as a tool for the measurement of implicit bias, however, it has yet to be replicated and was also constrained to very specific, contrived circumstances.

Additionally, Carpenter, Martinez, Vadhan, Barnes-Holmes, and Nunes (2012) employed implicit measures from the traditional IRAP pertaining to positive and negative consequences associated with cocaine use as a predictor of treatment attendance and subsequent cocaine use with cocaine-dependent participants. In addition to implicit IRAP measures, this study also incorporated another implicit measure, the Drug Stroop protocol and explicit survey measures, in order to determine which measure best predicted treatment outcomes. These authors found that the implicit measures obtained from the IRAP assessment, specifically those measures indicating derived relations pertaining to positive consequences of cocaine use and the absence of negative consequences of cocaine use, were moderately and significantly correlated with treatment outcomes. Essentially, this study demonstrated that the IRAP was able to moderately predict which participants would fare better or worse during treatment for cocaine dependence, based on their implicit attitudes towards the consequences of cocaine use, as measured by the IRAP.

The need for more research to address the predictive validity of IRAP and the fact that multiple instruments are now available with which to do so, set the basis for the following series of experiments. Specifically, the purpose of the first experiment of the present research was to conduct a direct comparison of the two iterations of the IRAP and determine the extent to which they produced convergent results on a within-subject basis. Additionally, to the extent they did not produce convergent results, an attempt to identify any sources of the divergence was conducted. Given that the first experiment of the present research could not answer the question of which IRAP iteration was “better” in any way (e.g., more accurate with respect to an individual’s patterns of overt behaviors of interest), the purpose of the second experiment of the present study was to determine the extent to which results of either instrument correlated with or predicted other patterns of behavior of interest across two groups of participants (Traditional IRAP group & Mixed Trial-IRAP group); specifically, in the case of the present study, patterns of behavior relating to work performance in an analog organizational setting. The results of the first and second experiments provided the impetus for the examination of the predictive utility of IRAP on a within-subject basis in the third experiment. In Experiment III, each participant completed both IRAP assessments (as in Experiment I), as well as the simulated work task (as in Experiment II), in order to determine which IRAP yielded greater predictive validity on a within-subject basis, using the same group of participants for each IRAP.

Experiment I

Method

Participants, Setting, and Apparatus

Participants in the present study were ten undergraduate students at the University of Nevada, Reno. Participants signed up through the Sona-system online subject pool management system, in which participants were instructed they must have at least two years cumulative work experience in either part-time or full-time job positions, in order to participate. Participants were offered extra credit for their time spent participating. The study took place in a small laboratory room in the psychology building on university campus. In the laboratory room was a standard PC desktop computer which ran the various computerized tasks with which participants interacted. There were two computer-based tasks that participants came into contact with; specifically, the two iterations of the IRAP (i.e., traditional IRAP and MT-IRAP).

Independent Variables

The independent variable manipulation for Experiment I consisted of simply exposing each participant to each of the two versions of the IRAP instrument, such that within-participant comparisons of the results could be made. Additionally, the sequence of exposure to each IRAP was counterbalanced across participants, in order to identify and control for any potential sequence effect.

Dependent Measures

The primary dependent measure was response latency on IRAP trials for each instrument. The raw latency data were transformed into D-IRAP scores, based on Greenwald et al.'s (1998) D-algorithm and Cohen's effect size *d* statistic. In addition to the response latency data and associated transformations, secondary dependent measures included error rates for both instruments.

Research Design and Procedure

The research design utilized in Experiment I was an AB/BA counterbalanced-across-participants exposure in which each condition represented exposure to a particular IRAP. Specifically, the A condition consisted of exposure to the traditional IRAP and the B condition consisted of exposure to the MT-IRAP. This sequence was counterbalanced across participants such that each successive participant was exposed to the sequence opposite that of the prior participant.

Upon arriving to participate in the study participants were first guided through the informed consent process, during which participants were explained general details of the study, and at which time they had the opportunity to ask questions about the study and consent to participate in the study, if they chose to do so. Following the informed consent process, participants were provided instructions for the particular IRAP iteration to which they were first exposed (see Appendix A for detailed instructions). In general, participants were given instructions which described the visual layout of the task on the computer screen and described how participants were to respond to the stimuli presented on the screen. Participants were also be shown a screenshot of the particular IRAP printed on a piece of paper, in order to help visualize the task and what was required of their responding. Participants were advised that they would be exposed to practice trials first to familiarize them with the task and that they had to meet certain performance criteria in order to pass the practice phases and move on to the test phases.

In the case of either iteration of the IRAP, following instruction regarding how to interact with the apparatus participants were then exposed to practice phases of the instrument, in order to become familiar with the response requirements of the instrument, as well as to determine whether the participant was able to interact appropriately with the instrument in order to continue on to the test phases of the assessment. If participants met the minimum performance criteria for inclusion in the study during the practice phases, those participants were then exposed to the test phases of each IRAP instrument, respectively. The following sections describe for each IRAP the nature of the practice phases, the specific performance criteria required to successfully complete the practice phases, and the subsequent test phases. The same evaluative and target stimuli, as well as relational response options, were utilized for both IRAPs, such that direct comparisons could be made, and these are shown in Table 1.

Traditional IRAP

Following instructions from the experimenter regarding how to interact with the assessment task, each participant was first exposed to a practice phase of the IRAP. The practice phase consisted of exposure to two consecutive blocks of trials, with each block requiring different patterns of responding on the part of the participant. Specifically, the first block required that participants respond in accordance with the first pattern of responding, which followed the rule of “like hard work words” and “dislike easy work words.” In practice, this required that participants respond “Yes” when presented with the stimulus combinations of Like→[Hard work word] and Dislike→[Easy work word] and that participants also responded “No” when presented with the combinations of Like→[Easy work word] and Dislike→[Hard work word]. In the event that a participant answered incorrectly (i.e., other than as specified above), feedback was provided in the form of a red “X” which appeared in the center of the screen, indicating an incorrect response, and the participant then had to emit the correct response in order for that trial to end and the subsequent trial to begin. Between each trial was an intertrial interval of 400 ms in which all stimuli were absent from the screen, and after which the next combination of stimuli for the following trial were then presented on screen.

Following completion of the first practice block participants were provided with feedback regarding their percentage accuracy during the block, in terms of correct trials out of total trials in the block, as well as the median response latency of all trials in the block. After the feedback was presented to participants on screen, the rule describing the required pattern of responding for the second practice block was presented to participants on screen. Specifically, participants were instructed to respond as though they “like easy work words” and “dislike hard work words.” As such, the correct responses were “No” in the presence of the stimulus combinations of Like→[Hard work word] and Dislike→[Easy work word] and “Yes” in the presence of the combinations of Dislike→[Hard work word] and Like→[Easy work word]. Again, if a participant responded incorrectly on any given trial, a red “X” appeared in the middle of the screen and the participant had to emit the correct response in order to proceed. As with the previous practice block, participants received performance feedback following completion of the block, pertaining to their accuracy and mean response latency.

For each practice block the performance criteria required to pass the block was a minimum of 70% accuracy and a median response latency of 2000 ms or less. Furthermore, in order to successfully to complete practice for the IRAP, these criteria had to be met for both practice blocks in succession. In other words, if a participant did not meet the criteria in the first block (i.e., first pattern of responding), but did meet the criteria for the second block (i.e., second pattern of responding), or vice versa, then this did not constitute successful completion of the practice phase. Each participant had a total of three opportunities to meet the pass criteria for both blocks consecutively. If a participant successfully completed the practice phase, then the participant moved on to the test phase of the IRAP.

For the traditional IRAP the test phase was identical to the practice phase with the exception that performance feedback (i.e., accuracy and average response latency) was not presented following each block. Participants completed three pairs of blocks, wherein each block within the pair required one of the two different response patterns, yielding a total of six blocks of trials. More specifically, the first, third, and fifth blocks of trials required participants to respond according to the first pattern of responses, which corresponded to the instruction of “respond as if you LIKE HARD WORK words and DISLIKE EASY WORK words.” The second, fourth, and sixth blocks required participants to respond according to the second pattern of responding, which corresponded to the instruction, “respond as if you LIKE EASY WORK words and DISLIKE HARD WORK words.” These instructions were presented to participants at the beginning of each block to which they corresponded, immediately prior to beginning the block of trials. Following completion of the sixth and final block of trials, the computer program advised participants that they were done with the IRAP and to inform the experimenter they were done.

In addition to the general procedures of the IRAP as outlined above, there were particular subtleties to the computer program itself that are pertinent to the present research and will therefore be discussed in the following portion of the procedure. In particular, the IRAP presented each target stimulus in combination with each of the two evaluative (i.e., “Like” or “Dislike”) stimuli in each block. Given that there were 16 target stimuli altogether (eight “hard work” and eight “easy work” words), there were exactly 32 trials per block. Given that there were six blocks in total, three corresponding to each pattern of responding, once the entire assessment was completed there were exactly three trials in which each participant responded “Yes” to a given combination of stimuli (e.g., Like→Stressful) and exactly three trials in which each participant responded “No” to that same combination. The traditional IRAP is programmed such that if a participant emitted an incorrect response (error) on a given trial, a red “X” appeared on the screen and the participant then had to emit the other, correct response before moving on. In such an instance, the IRAP program recorded the latency for the entire trial, beginning from the moment the stimuli were presented on screen (i.e., beginning of the trial), until the participant emitted the correct response, which necessarily included the time involved in emitting the incorrect response first and then emitting the subsequent correct response. When an error such as this occurred, the IRAP recorded the full trial duration (i.e., the latency until the final, correct response) as the datum for that particular trial. Therefore, for any given combination of stimuli (e.g., Like→Stressful), if a participant emitted an error on one of the trials, for example pressing the key for "Yes" instead of "No," then one of the three trials used for analyzing that particular stimulus combination response included the inflated latency that corresponded to the incorrect response followed by the correct response. (By the same logic, if two of the trials for a given response combination, for example, Like→Stressful→Yes were errors, then two of the three trials utilized for that analysis contained inflated error latencies, and so forth, such that all three latencies used for a particular stimulus combination response could have contained errors.)

In addition, while the latency criterion was set to 2000 ms per trial, after which time the "GO FASTER!" message appeared on screen, if a participant took longer than 2000 ms to respond (e.g., 5000 ms), the IRAP recorded that latency as the datum for the trial in question. These peculiarities of the traditional IRAP are at odds with some of the subtle procedural details of the MT-IRAP, discussed below, and should be noted for the purpose of comparing the two instruments.

MT-IRAP

Upon reading instructions for the MT-IRAP, participants were then exposed to practice phases of the MT-IRAP. Unlike the traditional IRAP, the MT-IRAP utilized different stimuli during the practice phases than during the test phases. Specifically, in the practice phase the MT-IRAP utilized a list of flower-type words (Rose, Flower, Tulip, Daisy) and a list of insect-type words (Beetle, Cockroach, Hornet, Spider). The two evaluative stimuli (Like and Dislike) remained consistent through both practice and test phases. In addition, the relational response options of "Yes" and "No" were consistent across practice and test phases. As noted previously, the MT-IRAP also incorporated a third stimulus on screen, specifically, the contextual cue consisting of either “Truth” or “Lie,” which were the same across practice and test phases.

During the first practice phase of the MT-IRAP, participants were instructed to respond according to the rule that they “Like flower-type words” and “Dislike insect-type words.” During the first practice phase, participants were only exposed to “Truth” trials, in which they had to answer “Yes” to stimulus combinations consisting of Like→[Flower word] and Dislike→[Insect word] and “No” to stimulus combinations consisting of Like→[Insect word] and Dislike→[Flower word]. In the event that a participant answered incorrectly (i.e., discordant with the instructions described above), then a red “X” appeared in the middle of the screen and the participant then had to emit the correct response to move on to the next trial. In addition, if a participant failed to respond within 3000 ms, then the message “GO FASTER!” appeared in the middle of the screen. Note that the response speed criterion for the MT-IRAP was 3000 ms, as opposed to 2000 ms for the traditional IRAP, given that the MT-IRAP required participants to respond to additional contextual cues (i.e., Truth or Lie) on each trial.

In order to successfully complete the first practice phase a participant had to achieve a minimum 70% accuracy on the trials within the practice block and an average response latency of 3000 ms or less. If a participant successfully completed the first practice phase, then the participant was done with that phase and was then exposed to the second practice phase of the MT-IRAP, in which all of the evaluative, target, and response stimuli were the same, with the addition of the “Lie” trials to the “Truth” trials. As previously noted, the purpose of the Truth/Lie stimuli was to occasion a participant’s emission of both response options (Yes or No) in the presence of each combination of evaluative and target stimuli. Therefore, in the case of the second practice phase, participants had to respond in the manner previously described for Truth trials and also had to reverse their answers on Lie trials. In other words, in the presence of the "Lie" cue, participants had to respond "Yes" to the stimulus combinations of Like→[Insect word] and Dislike→[Flower word] and "No" to the combinations of Like→[Flower word] and Dislike→[Insect word]. In order to successfully complete the second practice phase, participants had to respond with a minimum accuracy of 70% and maximum average response latency of 3000 ms for the entire block. Participants were allowed six total attempts to pass both blocks.

Upon successful completion of the practice phases, participants were then exposed to the test phase of the MT-IRAP. The test phase incorporated the same "hard work" and "easy work" target stimuli as used in the traditional IRAP (see Table 1). The evaluative stimuli, response options, and Truth/Lie cues were the same for the test phase as for the practice phases. Prior to beginning the test phase participants were advised that there were no longer "correct" or "incorrect" answers and that instead participants should respond to each combination of stimuli based on their own attitudes. For example, if a participant was presented with the stimulus combination of Like→Stressful→Truth, the participant was to respond however he or she saw fit, based on his or her own choice. Participants were instructed to respond based on their own attitudes, however, it should be noted that in the event a participant chose to answer according to socially appropriate or desirable responding, the participant may still have emitted such a response in the presence of the Truth label, even if that relational response did not cohere with the participant’s attitude toward the stimuli under other conditions in which the participant may be more “truthful” or “honest.” Even in such situations, the participant would still be required to emit the opposite relational response (Yes or No) in the presence of the Lie label, and ultimately the response latencies for those two response options were compared for the analysis, regardless of which of the Truth/Lie labels was present when the different responses (Yes or No) were emitted.

Additionally, participants were instructed to respond to each target stimulus individually and therefore not feel compelled to respond to one stimulus (e.g., Easy) in a particular way based on how the participant may have responded to a different target stimulus (e.g., Complex). In other words, it was possible for a participant to respond in terms of “Liking” work that is both “Easy” and “Complex” and that these target stimuli were not mutually exclusive, as they are treated in the traditional IRAP. Participants completed three blocks of trials in the MT-IRAP, in which both Truth and Lie trials were intermixed within each block. Once all three blocks were completed, the program advised participants they were done and to inform the experimenter.

As with the traditional IRAP it is important to specify some of the subtle procedural differences of the MT-IRAP, especially in terms of data collection. In particular, since there were no “right” or “wrong” responses in the test phase of the MT-IRAP, the program instead used the first block of trials to monitor and record each individual participant’s pattern of responding to each combination of stimuli. If during the first block a participant responded in a manner that was inconsistent for a given target stimulus, the computer presented the trials pertaining to that stimulus again sometime later in the block. The only assumption made for this purpose was that responses to a given combination of stimuli should be opposite in the presence of the Truth and Lie labels. For example, if in the presence of Like→Stressful→Truth a participant responded by pressing the key for "No," then in the presence of Like→Stressful→Lie the participant should have answered "Yes." If, however, a participant were to answer "No" to both combinations of stimuli (i.e., in the presence of both Truth and Lie labels), then the program recognized this as an inconsistent and invalid response pattern and cleared the responses of "No" to those stimuli as it tracked the participant's patterns of responding. In addition, the response latencies associated with those inconsistent and invalid responses were also cleared and not used in the final data analysis.

Instead, the program then re-presented those particular stimulus combinations in later trials within the same block and checked to see if the participant provided consistent (i.e., opposing response options in the presence of Truth and Lie) responses to those stimulus combinations. This procedure was repeated as many times as was necessary during the first test block for each different combination of stimuli, until the program had collected responses from the participant which were consistent for each stimulus combination. The final responses for each stimulus combination are those which were utilized for purposes of data analysis following the assessment. In conjunction with this procedural detail of the program, similar constraints were employed in the second and third test blocks. Specifically, if a participant responded during the second test block in a manner that was inconsistent with the recorded pattern of responses as determined during the first test block (e.g., responding "Yes" to Like→Stressful→Truth, after having responded "No" in the first block), then the program considered it an "incorrect" response, with respect to how the individual participant responded in the first block of trials. When this occurred, the program did not include the response latency for this "incorrect" (or inconsistent) response and instead re-presented that same trial again later in the same block, until the participant emitted a response to it that was "correct," or consistent with that participant's pattern of responding from the first block. This "errorless" process was employed in both the second and third blocks, such that once the assessment was completed, every response recorded for eventual analysis was "correct," in terms of being consistent with the participant's own uniquely established patterns of responding from the first block.

In addition to the "errorless" procedure described above, the MT-IRAP program also utilized a similar procedure for trials in which a participant emitted a response that took longer than the 3000 ms latency criterion. In particular, when a participant emitted a key-press response with a latency of greater than 3000 ms—regardless of whether the response was "correct" or "incorrect" (i.e., consistent or inconsistent with previous responding)—the MT-IRAP program did not use that latency in the final analysis and instead re-presented that same trial later during the same block, until the participant emitted a response to it that was *both* consistent *and* below the 3000 ms maximum latency criterion. In this way, the MT-IRAP ultimately produced a data set for final analysis that was devoid of any errors and trials that were answered more slowly than the criterion used to define and drive the implicit effect (3000 ms in this case). It should be further noted that all of the data for error trials and slow trials were saved and available for different analyses, but were not necessarily included in the data set used to calculate the IRAP effects.

In the event that a participant did not successfully complete the practice phases for either of the IRAP instruments, the participant was thanked for his or her time and excused from the study and the participant was provided with Psychology Experience Credits based on the standard rate of credits per time participated. Participants who completed both IRAP instruments successfully were thanked for their participation and provided with credits based on the standard rate per time participated. All participants were asked if they had any questions before they left and any questions participants had were answered at that time.

Data Preparation

The raw latency data for the IRAP instruments was transformed using a common algorithm, in order to analyze the results. The procedure is known as the D-algorithm and is similar to the transformation of data for implicit measures as initially employed by Greenwald et al. (1998). The procedure consisted of calculating a difference score between the mean response latencies of particular trials of interest and subsequently dividing that difference score by the standard deviation of all response latencies of interest. This transformation resulted in a D-IRAP score, which indicated a standardized difference in response latency for the responses in question. It has been assumed that the size of the D-IRAP score is an indication of the strength of the implicit attitude(s) being assessed. For the present research the D-IRAP transformation was completed at the level of lists of stimuli (i.e., “easy” and “hard” work words), as well as at the level of individual stimuli, for each individual participant, on a within-subject basis.

List-Level Analyses

The traditional IRAP outputs data in summarized form and provides researchers with a D-IRAP analysis at the level of the list of stimuli (e.g., “hard” work words), by Trial-Type. All published IRAP research to date, with the exception of the MT-IRAP research by Levin et al. (2010), has involved analyses conducted at the list level. Given this precedent, the first level of analysis for the present research was conducted at the level of lists of target stimuli. List-level analyses involve combining response latency data for all trials within a given Trial-Type. The four Trial-Types in the present research consisted of: 1) Like-Hard work, 2) Like-Easy work, 3) Dislike-Hard work, and 4) Dislike-Easy work. More specifically, in order to calculate the D-IRAP score for Trial-Type 1 (Like-Hard work), response latencies for all trials in which the evaluative stimulus at the top of the screen was “Like” and the target stimulus in the center of the screen was one of the eight “hard work” stimuli were included. According to the first response pattern required of participants in the IRAP, participants had to respond “Yes” when presented with “Like” and a “Hard work” word; this was the case in Blocks 1, 3, and 5 in the traditional IRAP. Conversely, according to the second response pattern required of participants in the traditional IRAP, participants had to answer “No” when presented with combinations consisting of “Like” and “Hard work” words; this was the case in Blocks 2, 4, and 6 in the traditional IRAP. In essence, for each pair of blocks (i.e., Blocks 1 & 2, Blocks 3 & 4, and Blocks 5 & 6), the trials in which “Yes” was the response were averaged to obtain a mean response latency for relationally responding “Yes” in the presence of Like→Hard work words (i.e., Trial-Type 1). The same was done for all trials in which “No” was the response for Trial-Type 1 trials.

The mean latency for all “Yes” trials was then subtracted from all “No” trials to yield a difference score between the two means; the direction of this arithmetical procedure resulted in a positive score indicating a Pro-Hard work bias (i.e., responding “Yes” more quickly to Like→Hard work words than “No”), and likewise, a negative difference score indicated a Pro-Easy work bias (i.e., responding “No” more quickly to Like→Hard work words than “Yes”). The difference in score was then divided by the standard deviation of all the response latencies which were included in the two means (i.e., all Trial-Type 1 trials for a given pair of blocks). This procedure yielded the D-IRAP score for Trial-Type 1 trials for each of the three pairs of blocks. Those three D-IRAP scores were then averaged to yield a total D-IRAP score for Trial-Type 1, which is conceptualized as a measure of the strength of the implicit attitudes (i.e., relational responding) with respect to the stimuli in question (i.e., “Like” and “Hard work” words). The equations below represent this procedure arithmetically:

$$\text{D-IRAP Trial-Type } 1_{B1} \quad = \quad \frac{M(Yes_{B1}) - M(No_{B1})}{SD(All_{B1})}$$

where B indicates the pair of test blocks included in the analysis (in this case, test Blocks 1 & 2 constitute the first pair of blocks used to calculate the first D-IRAP score for Trial-Type 1), Yes refers to response latencies for relational responses of “Yes,” No refers to response latencies for relational responses of “No,” and All refers to all response latencies for the pair of blocks. The above calculation was conducted for the three pairs of test blocks (i.e., $_{B1}$ = Blocks 1 & 2, $_{B2}$ = Blocks 3 & 4, $_{B3}$ = Blocks 5 & 6), and then the resulting three D-IRAP scores were averaged to yield a single D-IRAP Score for Trial-Type 1. Note that for Trial-Types 2 and 3, the numerator was reversed, such that the mean of "No" latencies was subtracted from the mean of "Yes" latencies (additionally, one could use the equation as written above and simply multiply the result by -1), in order to maintain consistency with a positive score indicating a Pro-Hard work bias and a negative score indicating a Pro-Easy work bias.

The procedure described above was then repeated for the other three Trial-Types, which yielded a total of four D-IRAP scores, one for each of the four different Trial-Types. These Trial-Types were parsed in this way because RFT does not assume that Like→Hard work is functionally equivalent to Dislike→Easy work, or that Like→Hard work is functionally opposite of Dislike→Hard work. Instead, each Trial-Type represents a specific relation among the particular stimuli in question, which may be functionally independent of any of the other specific relations being assessed. Furthermore, insofar as the IRAP is a measure of an individual's most probable brief, immediate relational responses to certain stimuli, which are influenced by that individual's learning history and history of relational responding to those stimuli, it is possible that an individual's behavioral repertoire consists of much more fluent, practiced relational responses to some of the four Trial-Types than to others. In other words, an individual may have a much more extensive history of relational responding in terms of "Liking" certain things or events (e.g., "Simple" work or "Relaxing") than "Disliking" those things. Nevertheless, in previous IRAP research (see Barnes-Holmes et al., 2010) all four D-IRAP scores have been averaged to yield a single, overall D-IRAP score, which has been referred to as a measure of a general IRAP Effect. The extent to which this single score is positive or negative is said to indicate, on the whole, an individual's pre-existing bias with respect to the attitudes in question, in this case, either a Pro-Hard work or Pro-Easy work bias, respectively.

The list-level analysis was also conducted with the MT-IRAP response latency data in the same manner as described above for the traditional IRAP. For clarification purposes, the Trial-Type 1 latencies were those in the presence of "Like" and "Hard work" words. In the presence of the "Truth" label, a participant responded either "Yes" or "No" to these trials, and in the presence of the "Lie" label, a participant responded in the opposite manner, thus yielding response latencies of both "Yes" and "No" for the particular Trial-Type 1 combinations. The D-IRAP score analysis then proceeded exactly as outlined above for all four Trial-Types.

There are important, inherent procedural differences between the two iterations of the IRAP instrument which are of interest when calculating the D-IRAP scores for each instrument. In particular, the traditional IRAP is programmed to remove any response latencies over 10,000 ms. There were no latencies removed from any participant data set on this basis. The traditional IRAP also claims to automatically round any response latencies between 3000 ms and 10,000 ms down to 3000 ms; however, upon close inspection of the analyses produced by the IRAP, it was found that response latencies greater than 3000 ms were left as such, rather than being rounded down to 3000 ms. Additionally, according to the current precedent in research with the traditional IRAP, a participant's data set was removed entirely if the participant responded to more than 10% of trials with a response latency of 300 ms or less; no participants' data were excluded on this basis. Lastly, according to recent research with the traditional IRAP and the only study to date to investigate the predictive validity of the traditional IRAP (Nicholson & Barnes-Holmes, 2012), a participant's data was excluded if the participant responded with less than 75% accuracy for any two of the six test blocks; no participant's data were removed on this basis. Similarly, if a participant responded with less than 75% accuracy for one test block, then that block and its corresponding (i.e., Pattern 1 or Pattern 2) test block were removed from the analysis, while the other blocks of trials remained; one participant's (PP1-9) data for the second pair of test blocks were removed on this basis.

There are idiosyncratic differences in data preparation for the MT-IRAP as well. In particular, and as noted previously, the MT-IRAP does not include response latencies on error trials in the final analysis. Instead, the program re-presented the trial on which an error was emitted until a correct response was emitted and its latency recorded. Similarly, if a response latency for a specific trial exceeded the 3000 ms time limit, after which point the "GO FASTER!" message appeared, the program re-presented that trial until a latency for a correct response which was less than 3000 ms could be recorded.

One potential drawback of the procedural details of the MT-IRAP, particularly pertaining to the aspect whereby participants choose how they respond to each combination of evaluative and target stimuli (as opposed to being forced to respond one way or another), is the fact that how participants respond during the first test block, in which the program determines what is consistent or "correct" for each unique participant, may not be how participants respond in later test blocks. In other words, participants' responding may change from the first test block to successive test blocks. There appear to be two primary reasons such a shift in responding may occur: one is that participants may simply "change their minds" about their explicit attitudes regarding the stimuli in question, and therefore begin to emit the other relational response option on later trials in later blocks; the other potential source for this shift in responding could be that participants accidentally emit the “incorrect” relational response (i.e., make a mistake), with respect to their own choices regarding the stimuli in question, consistently enough during the first test block that the program records and determines those “incorrect” responses as being the “correct” responses for those participants for the stimuli in question. In such a scenario, if a participant were to then emit what that participant chooses to be the “correct” response to the stimuli in question in blocks two or three, the program would recognize these latter, “correct” responses as “incorrect.”

When a shift in responding as described above occurs for any reason, it tends to result in the participant being exposed to the same combination of stimuli (e.g., Like→Stressful→Lie) repeatedly as the participant nears the end of the block and the program re-presents only those stimulus combinations which do not yet have a fast (i.e., 3000 ms or less) and accurate (i.e., “correct”) response recorded for that block. This perpetual loop of a specific trial is typically eventually terminated when the participant emits the other, “incorrect” response relative to what has been repeatedly emitted during the loop, which the program then recognizes as “correct,” thereby ending the block.

In the present analysis, the response latency data for any stimulus for which this occurred were removed from the list-level analyses for a given participant. Partial data for six participants were excluded on this basis in the present study. Specifically, Calm, Simple, Demanding, and Relaxing were removed for participant PP1-6; Calm, Simple, and Leisurely were removed for participant PP1-7; Slow, Demanding, and Persistence were removed for participant PP1-8; Calm, Slow, Relaxing, Lazy, Quitting, and Stressful were removed for participant PP1-9; Simple and Difficult were removed for participant PP1-11; and Hard, Difficult, Busy, and Demanding were removed for participant PP1-12.

Results

Figures 4 and 5 depict list-level analyses, inclusive of all four Trial Types and the overall IRAP effect, for two representative participants in Experiment 1. The results shown in Figure 4 (PP1-1) were representative of a strong, positive correlation among the two IRAP instruments on a within-subject basis. Figure 5 depicts a participant whose data indicated a weak negative correlation among the two instruments (PP1-11). In each figure, percentage correct for each of the four Trial Types for traditional IRAP only is shown in red. There did not appear to be any sequence effects as a result of sequence of exposure to the two iterations of the instrument, based on visual inspection of the data.

Generally, it was found that the two IRAP instruments provided largely divergent data for half of participants, while there was some degree of convergence for the other half of participants, regardless of which IRAP was conducted first. Although RFT and thus the IRAP do not assume any functional dependence among the four Trial Types, broad divergence among several of the Trial Types raises suspicions about the internal consistency of the assessment. Such divergence at the list-level of analysis was observed with the traditional IRAP for a number of participants; specifically, participants PP1-1, PP1-2, PP1-5, PP1-7, PP1-8, PP1-9, and PP1-11, which constituted 70% of participants. Similar divergence among Trial Types within a given IRAP were also observed with the MT-IRAP; specifically, among participants PP1-2, PP1-9, and PP1-11, which constituted 30% of participants. In these instances in which divergent D-IRAP scores were obtained across the Trial Types, the overall IRAP Effect score tended to be closer to zero, as a result of the divergent scores cancelling each other out when averaged.

At surface level, the MT-IRAP appeared to produce results that were more consistent on a within-subject basis. For example, although not all Trial Type scores were necessarily in the same direction (i.e., indicative of a similar bias) for all MT-IRAP assessments, a greater majority of Trial Type scores were in a similar direction (i.e., positive or negative D-IRAP scores), indicating a similar direction of preference or bias with respect to the particular attitudes being assessed. Although it is not possible to say which IRAP instrument produced results that were closer to any kind of “true” target, it can be definitively said that the two instruments did not produce consistently convergent, reliably similar results, regardless of whichever is more “accurate.”

Correlational analyses were conducted to determine the extent to which the two IRAP instruments produced convergent data sets and are presented in Table 2. For each individual participant, Pearson’s product-moment correlation coefficient *r* was calculated between the traditional IRAP and MT-IRAP for each of the four Trial Types. Moderate to strong positive correlations were observed for half of participants, particularly PP1-1 (*r* = .85), PP1-6 (*r* = .59), PP1-7 (*r* = .64), PP1-8 (*r* = .92), and PP1-9 (*r* = .63). A weak positive correlation was observed for PP1-2 (*r* = .29). Data sets for several participants indicated weak positive or negative correlations (PP1-4, *r* = -.11; PP1-5, *r* = -.18; PP1-11, *r* = -.16; and PP1-12, *r* = .18). While the correlation coefficients for some individual participants were fairly strong and positive, the overall correlation coefficient for all Trial Types across all participants combined was positive and fairly weak (*r* = .19) and was not statistically significant (directional *p* = .12, non-directional *p* = .23).

In addition, when all four Trial Types were averaged into the single IRAP Effect for each participant for each IRAP instrument and those IRAP Effects were then subjected to correlational analysis, the result was a weak-to-moderate, negative correlation (*r* = -.34) among the two instruments. It appears that this negative correlation is due at least in part to the fact that the overall IRAP Effect for the traditional IRAP data tended to be closer to zero (six participants' overall IRAP Effects were less than 0.1 different from zero, and the other four participants' IRAP Effects were less than 0.2 different from zero), as a result of the disparate Trial Types being averaged together. Additionally, the overall IRAP Effects also tended to be in the opposite direction of the overall IRAP Effect for the MT-IRAP data (eight of 10 participants' overall IRAP Effects were in differing directions across the two IRAPs), while the MT-IRAP tended to be less disparate across Trial Types for each participant. When the absolute values of the overall IRAP Effects for each IRAP instrument were averaged across participants, the average for the traditional IRAP was 0.087, while the average for the MT-IRAP was 0.296, suggesting that the MT-IRAP produced more consistent Trial Type D-IRAP scores, in terms of direction (i.e., positive indicating Pro-hard work bias, negative indicating Pro-easy work bias), which were averaged into the overall IRAP Effect scores.

Since the overall IRAP Effect scores were affected in the manner described above, and again recognizing that distinct Trial Types are not inherently functionally related, nor are participants likely to be as fluent in relational responses associated with each Trial Type, subsequent analyses were conducted using the four distinct Trial Types, rather than averaging them into a single IRAP Effect score. As such, correlational analyses were also conducted using the D-IRAP scores for each Trial-Type for all participants, across both IRAP instruments. The correlation coefficient for both IRAPs, including all participants, for Trial Type 1 was *r* = .73; Trial Type 2, *r* = -.31; Trial Type 3, *r* = -.55; and Trial Type 4, *r* = .30.

Given the large discrepancies in D-IRAP scores across the different Trial Types per participant, especially with the traditional IRAP, an attempt was made to identify a possible source of the disparities. In particular, the response latencies associated with error trials were removed from the analyses, in order to determine whether the internal consistency (in terms of D-IRAP scores for each Trial Type) for the traditional IRAP would improve, and also to determine whether the correlations with MT-IRAP scores would improve. Figures 6 and 7 display the same analyses as presented in Figures 4 and 5, however, the response latencies for all error trials in the traditional IRAP have been removed.

Correlational analyses were conducted between data sets for the two IRAP instruments, in order to determine whether the response latencies associated with errors from the traditional IRAP appeared to have influenced the disparities noted within each traditional IRAP assessment and as compared with MT-IRAP assessments. The results of the correlational analyses are presented in Table 3. The correlation coefficients across Trial-Types for six individual participants improved somewhat, while for the other four the correlation coefficient moved in the opposite direction (i.e., less positively or more negatively correlated). The correlation coefficient for all Trial-Types across all participants improved only slightly, from *r* = .19 to .23. Additionally, when the overall IRAP Effect was calculated for each IRAP for each participant, the correlation coefficient among the two IRAPs improved markedly from *r* = -.34 to -.08, once errors trials were removed from the traditional IRAP. In conjunction with the somewhat improved correlations among the two instruments once errors were removed, the average absolute value of each overall IRAP Effect for all participants for the traditional IRAP increased from 0.087 to 0.109, indicating that the D-IRAP scores for the individual Trial-Types for each participant's IRAP were more consistent, in terms of the direction (i.e., bias) of the scores.

In addition, the correlation coefficients for each Trial-Type, across all participants, were reanalyzed with the removal of error latencies. For Trial Type 1, the correlation coefficient decreased slightly, from r = .73 to .68. The corresponding error rates for Trial Type 1 were relatively low, with all participants achieving an average accuracy of 93.8% correct on Trial Type 1 trials. For Trial Type 2, the correlation coefficient improved slightly, from r = -.31 to -.26; the initial average accuracy across participants for Trial Type 2 was also 93.8%. The correlation coefficient for Trial Type 3 trials also improved slightly when the error latencies were removed from the analysis, from r = -.55 to -.51; the initial accuracy for Trial Type 3 was 87.3%. Lastly, the correlation coefficient for Trial Type 4 improved dramatically, from r = .30 to .74, which is a strong positive correlation. Interestingly, the initial accuracy across all participants for Trial Type 4 was 76.1% and it was by far the Trial Type with which most participants struggled, in terms of accurate responding. In general, it was observed that the "Dislike" Trial Types (i.e., Trial Types 3 & 4) were those to which participants responded inaccurately most often.

The frequency of errors across the different Trial Types was analyzed further to determine whether there were significant and meaningful differences in the frequency errors observed. The following means and variances for errors for each Trial Type in the traditional IRAP were subjected to a single-factor ANOVA: Trial Type 1 M = 2.5, s^2 = 2.94; Trial Type 2 M = 2.4, s^2= 2.04; Trial Type 3 M = 5.1, s^2= 13.21; Trial Type 4 M = 8.9, s^2= 18.32. The resulting ANOVA found F(3, 36) = 10.2, p = .0000. A similar analysis was conducted for the MT-IRAP errors across different Trial Types, with means and variances of: Trial Type 1 M = 5.3, s^2= 35.57; Trial Type 2 M = 6.5, s^2= 43.83; Trial Type 3 M = 15.9, s^2= 132.10; Trial Type 4 M = 14.4, s^2= 96.93. The ANOVA analysis yielded F(3, 36) = 3.78, p = .01. These analyses indicated that the "Dislike" Trial Types were more difficult to respond to and occasioned significantly greater frequencies of errors on the part of participants.

Stimulus-Level Analyses

In addition to the list-level analyses presented above, analyses were conducted at the level of the individual stimulus, since there is no reason to assume that the various stimuli categorized into each list are functionally similar, in terms of their psychological functions for different participants. The raw response latency data were again transformed into D-IRAP scores for both the traditional IRAP and MT-IRAP data, however, the data transformation process was slightly different from that employed in the list-level analyses. Specifically, and based on the procedures described in Levin et al. (2010), the D-IRAP scores for each target stimulus were calculated using the response latencies for that target stimulus and each of the separate evaluative stimuli (i.e., Like and Dislike), across all test blocks of trials. For example, for the combination of Like→Effortful, the three response latencies for the "Yes" relational response and the three latencies for the "No" relational response, which were drawn from all three of the test blocks, were utilized to calculate the mean latency for the “Yes” and “No” responses, respectively. From there, the pooled standard deviation of those latencies was calculated from those trials combined (i.e., Yes and No responses), and the difference between the two means was divided by that pooled standard deviation. This in turn yielded a D-IRAP score for the combination of Like→Effortful. The same procedure was repeated for the combination of Dislike→Effortful, to yield a D-IRAP score specific to that stimulus combination. This was done for the same reason as the individual Trial-Types were parsed out in the above analysis—namely, avoiding any assumption that the combination of Like-Effortful is in any way functionally related to Dislike-Effortful. This was analogous to the four different Trial-Types in the list-level analyses, however, since this analysis proceeded at the level of each individual stimulus, there were essentially only two Trial-Types per stimulus: that of “Like” and the stimulus in question, and “Dislike” and the stimulus in question. The data transformation procedure is represented below arithmetically:

$$\text{MT-IRAP}_{E1T1} = \frac{M(\text{Yes}_{E1T1}) - M(\text{No}_{E1T1})}{\text{Pooled SD}_{E1T1}}$$

and

$$\text{Pooled SD}_{E1T1} = \sqrt{\underline{(\text{SDYes}_{E1T1})^2(\text{nYes}_{E1T1} - 1) + (\text{SDNo}_{E1T1})^2(\text{nNo}_{E1T1} - 1)}}$$

$$(nYes_{E1T1} - 1) + (nNo_{E1T1} - 1)$$

where E is one of the two evaluative stimuli (Like or Dislike), T is one of the 16 individual target stimuli (e.g., Stressful), Yes refers to response latencies for the relational response of Yes, and No refers to response latencies for the relational response of No.

Figures 8 and 9 depict the D-IRAP scores for each stimulus combination from each of the two IRAP instruments (traditional IRAP in top panel, MT-IRAP in bottom panel), for two representative participants. Figure 8 shows PP1-9, who exhibited the strongest positive correlation of *r* = .67, and Figure 9 depicts results for PP1-2, who produced the strongest negative correlation of *r* = -.21. The percentage of correct trials for each stimulus combination in the traditional IRAP are displayed in red in each figure. The errors were not removed for this analysis, since there were so few data to begin with when analyzing each stimulus combination (i.e., six latency data per combination, for example, Like→Stressful, as described above in the data transformation procedure). Any stimuli that were removed from the list-level analyses for the MT-IRAP, as described previously, were left in for these analyses and are indicated as such in the figures.

Once again, as with the list-level analyses, there were largely disparate results from one IRAP instrument to the next, when measured at the level of the individual stimulus. Table 4 presents the Pearson product-moment correlation coefficient *r* for each participant, based on the D-IRAP scores for all of the individual stimuli, of both IRAP instruments. The range of correlation coefficients was -.21 to .67. While data for only three participants exhibited a negative correlation coefficient (PP1-2, PP1-5, PP1-11), only three participants' data indicated moderate-to-strong positive correlational relationships (PP1-7, PP1-8, PP1-9), with one of them being fairly weak (i.e., PP1-8, *r* = .25). When all participants' data for all individual stimuli were entered into a single Pearson's analysis, the coefficient was *r* = .15, again indicating a weak positive correlation between the two IRAP instruments, which was roughly identical to the correlation observed when comparing the data at the level of lists of stimuli and Trial-Types. However, given that there were so many data entered into this correlational analysis, the correlation was found to be statistically significant at *p* = .01. Pearson's coefficients were also calculated for all participants across the two "trial types"; in other words, D-IRAP scores for each of the stimuli paired with the "Like" evaluative stimulus were conducted, as were correlations for stimuli paired with the "Dislike" evaluative stimulus. These coefficients are also presented in Table 4. In general, only a few moderate-to-strong correlations were observed for either trial type (i.e., Like or Dislike).

Discussion

Experiment I of the present study sought to determine whether the two iterations of the IRAP instrument currently in use and available for research—the traditional IRAP and the MT-IRAP—would produce similar, convergent results when administered within-subject and with little or no time between administrations (in order to control for any extraneous factors which may influence implicit responding). In general, it was found that the two instruments did not yield similar results when compared within-subject at the level of individual stimuli, at the level of lists of stimuli, or when a single, overall

IRAP Effect score was calculated. There were few instances in which moderate or strong positive correlations were observed within subject across the two instruments, however, there were greater instances in which no correlation or even negative correlations were observed within subject. Additionally, when all participants' data were grouped together, very weak positive correlations were observed at both the individual stimuli- and list-levels of analysis, however, a moderate negative correlation was observed at the level of the single, overall IRAP Effect.

It is difficult to determine exactly which factors contributed to the observed discrepancies between the two instruments without designing a study to specifically address that question; however, based on our initial analyses of the instruments and respective data, we offer some suggestions. Firstly, as indicated in the procedural and results sections above, the two instruments have fairly different ways of collecting data for purposes of analysis. More specifically, the traditional IRAP presents a fixed number of trials per block, in which each combination of evaluative stimulus and target stimulus is presented only once. As such, the response latency recorded for each trial is kept for purposes of analysis, even if the response on a given trial is incorrect or well beyond the time limit imposed in the implicit measure (or both). Additionally, it is worth noting here that the traditional IRAP claims to take response latencies that are greater than 3000 ms and less than 10,000 ms and round them down to 3000 ms; while this rounding can be observed in some of the data output files of the traditional IRAP, it does not actually occur in the analyses in the most commonly used output file, which provides researchers with pre-determined D-IRAP Scores for each Trial-Type at the list level of analysis.

Conversely, the MT-IRAP uses different procedures to collect participant data and determine which data are utilized for purposes of analysis. In particular, the MT-IRAP presents every possible combination of evaluative and target stimuli at least once per block, but also presents some of those stimulus combinations multiple times within a given block, if necessary, based on participant responding. In the event that a participant responds to a certain trial in an incorrect manner, the MT-IRAP presents that same combination again sometime later within the same block, in order to provide the participant another opportunity to respond correctly to the stimulus combination. Similarly, if a participant provides a response to a given trial that is beyond the time limit set by the experimenter as the latency limit to drive implicit responding, the MT-IRAP will present that same stimulus combination again later within the same block, such that the participant will have another opportunity to respond both quickly (i.e., within the implicit response latency limit) and accurately. Ultimately, it is these fast and accurate responses which are used for purposes of analysis, although the other slow and incorrect responses are still recorded in raw data files, in case they are needed for other analyses.

It is difficult to say precisely to what extent these differences in the two instruments account for the observed differences in their results, given that we cannot know with what to replace the slow and incorrect response data from the traditional IRAP. However, in order to attempt to clarify the relationship of this procedural difference with the disparate results as much as possible, analyses were conducted in which the incorrect response latencies were removed from the data sets for the traditional IRAP. In general, doing so improved the overall correlation between the two instruments at the level of lists of stimuli slightly, and had an even greater impact on the correlation between the two instruments at the level of the overall IRAP Effect. In particular, analyses conducted at the list level with each Trial Type indicated that the largest improvement in correlation coefficients, in terms of producing a stronger positive correlation, was observed for Trial Type 4 (Dislike-Easy work words), which was the single Trial Type with the greatest amount of errors for virtually all participants. Taken together, these analyses suggest that at least some amount of the discrepancy in results between the two IRAP instruments is attributable to these procedural differences among the instruments, in terms of how they collect and utilize participant response latencies.

While the data sets for the MT-IRAP appear, at surface level, to have greater face validity, in terms of D-IRAP scores for the various Trial Types being more consistent in their direction of implicit bias (i.e., most or all Trial Type scores indicating a similar Pro-Easy work or Pro-Hard work bias), it cannot be said with certainty that the MT-IRAP measures are any closer to a participant's "true" implicit attitudes or beliefs. In fact, these two instruments are the only ones presently available to behavioral researchers to measure implicit responding, and to the extent that their measures are not convergent, there is no additional way to determine which is more accurate with respect to an individual's "true" or actual implicit attitudes. Nonetheless, it would appear at this time, based on the factors elucidated above, that the MT-IRAP probably provides a more consistent, precise measure of implicit attitudes relative to the traditional IRAP. At this point in time, however, additional research is needed to support or refute this assertion.

One potential means of attempting to validate each IRAP instrument is to determine the extent to which the results of either instrument correlate with or predict some other overt pattern of behavior. Such research has been discussed conceptually as important to the field of implicit cognition research (De Houwer, 2002; Nicholson & Barnes-Holmes, 2012), however, little research of the sort utilizing the IRAP has been conducted to date. It remains an open question whether or not implicit attitudes predict, correlate with, or to some extent influence other patterns of behavior which occur in naturalistic settings and with which researchers and practitioners are interested. Using such experimental paradigms to inform the question of which IRAP instrument produces more “accurate” results with respect to an individual’s implicit attitudes rests on the assumption that there is at least some relationship between implicit attitudes as measured by the IRAP and other overt patterns of behavior. While making such assumptions is perhaps not ideal, it seems necessary to move forward with investigations aimed at comparing the two IRAP instruments, in the first place, and further determining which, if any, correlates with or predicts other patterns of behavior with which we are more interested, particularly for applied purposes.

The argument stated above formed the basis for the second experiment within this study, the purpose of which was to determine the extent to which either IRAP instrument produced results which correlated with or predicted other patterns of overt behavior. The overt patterns to be correlated with IRAP results consisted of participant performance on a pay-for-performance, simulated data entry work task, in an analog organizational setting.

Experiment II

Method

Participants, Setting, and Apparatus

Participants in Experiment II were five undergraduate students at the University of Nevada, Reno, who did not participate in Experiment I. Participants signed up through the Sona-system online subject pool management system, in which participants were instructed they must have at least two years cumulative work experience in either part-time or full-time job positions, in order to participate. The study took place in the same small laboratory rooms on university campus in which Experiment I took place. Similarly, the same PC desktop computers were utilized to conduct Experiment II and they also had the simulated data entry work task, in addition to the two versions of the IRAP instrument.

Independent Variables

The independent variables for Experiment II were the IRAP instruments to which participants were exposed and the increasing work demands placed on participants during the simulated data entry work task. More specifically, during the simulated work task participants had to perform at an increasingly demanding minimum level of performance, in terms of correct trials per 2-min session. Periodically, the work demands per session increased (described in detail below).

Dependent Measures

The primary dependent measures were the response latencies as measured by the IRAP instruments (as in Experiment I) and the performance measures associated with the work task. Specifically, primary measures for the work task consisted of number of 2-min work sessions completed; average number of correct trials per 2-min session; and persistence, in terms of number of sessions completed during the work task. Secondary dependent measures consisted of Likert-type self-report ratings of how stressful and demanding the work task was at different points in time throughout the task.

Research Design and Procedure

The research design was a between-groups analysis of the extent to which IRAP scores from the different versions of the IRAP correlated with the dependent measures from the simulated work task. More specifically, one group of participants was exposed to only the traditional IRAP and the simulated work task, while the other group of participants was exposed to only the MT-IRAP and the simulated work task. The IRAP results for each group were correlated with the work task performance of each group, in order to determine which IRAP produced results that better correlated with work task measures, at the group level. Two participants first completed the simulated data entry work task, followed by one of the two IRAP instruments. The other three participants completed the IRAP assessment first, followed by the work task. The counterbalancing of the work task with the IRAP assessments served not only to control for any potential sequence effects (e.g., priming effects concerning persisting at or quitting the work task, based on exposure to those terms during an IRAP), but also to determine whether completing the work task under conditions of frustration and fatigue (i.e., following the an IRAP assessment, which participants have anecdotally reported can be frustrating and "cognitively taxing") influences the extent to which implicit attitudes as measured by the IRAP correlate with or predict behavior with respect to the work task.

Upon arriving to participate in the study participants were first guided through the informed consent process, during which participants were explained general details of the study, and at which time they had the opportunity to ask questions about the study and consent to participate in the study, if they chose to do so. Following the informed consent process, participants were provided instructions for the particular IRAP iteration to which they were exposed (see Appendix A for detailed instructions). All aspects of the procedure pertaining to the IRAP instruments were identical to those described for Experiment I.

Participants completed either the IRAP assessment or the simulated work task based on which sequence of exposures to which they were assigned. When it was time to engage with the data entry task, participants were shown the work task and worked through some sample trials with the experimenter. Following the sample trials, the experimenter then left the room and allowed participants to engage with the work task. The specific procedural details of the simulated work task are described in detail below (refer to Figure 10 for a screenshot of the task).

The simulated work task was a data entry task based loosely on the job of an entry-level EKG technician. In general, at the beginning of each trial, information for a hypothetical patient appeared in the upper-left corner of the workspace screen. The participant was required to reference the medical information provided for the particular patient against charted specifications found in the upper-right and right-hand portions of the work screen, in order to determine if the patient’s medical information was below, within, or above certain parameters for each medical metric. In particular, the participant had to determine if the patient’s “QT Interval” was below, within, or above the range specified based on patient gender, and also whether the patient’s heart rate (HR) was below average, average, or above average, based on the average HRs for certain age groups (regardless of patient gender). The responses were made using the mouse by checking one of the radio buttons that corresponded to the different ranges of below, within/average, or above.

Each 2-min work session began when a participant clicked the button on the screen which read “Start next session.” Once it was clicked, a timer began to count down from 120 seconds and the first patient’s data appeared in the upper-left portion of the work screen. Once the participant checked the appropriate radio buttons based on the patient information, the participant then clicked the “Submit” button. Upon clicking the Submit button, the Submit button disappeared, as did the medical information pertaining to the patient. If the choices (i.e., radio buttons selected) made by the participant were correct, then the revenue counter at the bottom of the work screen flashed green for one second and counted upward by three cents. If the choices made by the participant were incorrect, then a large red “X” appeared where the Submit button was for one second. Between each trial was an intertrial-interval of 500 ms, during which there was no patient information, no Submit button, no red “X” and no green coloration of the revenue counter. Following the intertrial-interval, the medical information for the next patient appeared, as did the Submit button, and the participant was then able to complete the next trial. If a participant clicked the Submit button while having selected only one or none of the radio buttons, then the trial was automatically considered incorrect.

As the participant completed correct trials the participant accumulated revenue earned, as tracked on screen throughout the task, which was paid in cash to the participant at the close of the experimental session. The money earned and paid out to participants was based on performance during each 2-min session. Specifically, as previously noted, there was a minimum work requirement, in terms of correct trials completed, for each 2-min session. In order to keep the money earned during a given 2-min session and have it accumulate toward the final payout, the participant had to meet the minimum work requirement for that 2-min session. In the event that a participant did not meet the minimum work requirement for a given 2-minute session, the money earned during that session was not added to the cumulative total to be paid out. However, any money that was earned during prior 2-min sessions, based on meeting the minimum work requirements in those sessions, was not removed. As such, the minimum requirement to accumulate money to be paid out was on a session-by-session basis.

At the close of each 2-min session, participants were presented with feedback regarding their performance in the most recent session (see Figure 11). The feedback informed participants of whether they did or did not meet the minimum work requirement for that session, and whether they did or did not get to keep the money earned during that session, based on having met the work requirement or not. The feedback also informed participants of how much money from that session was added to the cumulative total, as well as how many correct trials they answered out of trials attempted. Participants then clicked on the “Continue” button to exit the feedback screen and return to the work task screen to begin the subsequent 2-min work session.

The minimum work requirement for the first three 2-min sessions was eight correct trials per session. Following the completion of the first three sessions, the minimum requirement for the next three sessions increased by two, up to 10 correct trials per 2-min session. Every three sessions the work requirement increased by two correct trials per session. The increase in work requirements and timing of the increases was held constant across all participants, such that the quantifiable aspects of the “work demand," in terms of correct trials required, was held constant and was not a source of potential confound, even though different participants responded at different rates (and therefore failed to be able to meet the minimum requirements at different rates).

In addition to the increasing work requirements, participants were periodically asked to rate on a scale of one to seven how stressed they currently felt, followed by how demanding they currently found the work task. These ratings questions appeared every six sessions and appeared on the computer screen in between the 2-min work sessions, after the feedback screen from the prior session and before returning to the work screen for the next session. Additionally, participants were also presented with the stress and demand scale questions immediately following any session in which they did not meet the minimum work requirements for that session.

One of the most important features of the simulated work task was the "Quit" button, which was located in the work space screen and was present at all times (except during the feedback and stress/demand screens, during which the work screen was not visible). Participants were instructed prior to beginning the task that they had the option to quit at any time, by clicking the Quit button, and they would be paid whatever they had earned up to the point they chose to quit, based on cumulative revenue successfully earned. Upon clicking the Quit button, participants were asked if they really wanted to quit, in order to avoid participants accidentally quitting the task prematurely, and if they responded "Yes," then the computer program advised the participants that the task was over and to please notify the experimenter. At that time, participants were paid in cash whatever they had successfully earned in revenue throughout the work task and were also asked to complete a couple of questionnaires.

The first questionnaire was a survey of work attitudes, corresponding to the target stimuli utilized in the IRAP assessments, presented in question form and to which participants could take as much time to answer as they pleased (i.e., explicit survey of work attitudes; see Appendix B). Following completion of the explicit work attitudes survey, participants were then asked to complete a short questionnaire regarding their experiences during the experiment (see Appendix C). Questions asked participants about their primary motivation for participating in the study (e.g., extra credit, money, both), how stressful and demanding they found the work task to be, and their motivation for deciding to quit the work task when they did.

Results

Primary dependent measures collected for each participant included IRAP measures for each participant, which were analyzed at both the list and individual stimulus levels. Some participants were exposed to only the MT-IRAP instrument (PP2-1 and PP2-3), and others were exposed to only the traditional IRAP (PP2-2, PP2-4, and PP2-5). In addition to IRAP measures, performance data on the simulated data entry task were collected for each participant. Work task performance measures are presented for all participants in Table 5.

In order to determine the extent to which a given participant's IRAP results correlated with or predicted performance on the data entry task, a number of comparisons were made across a number of different metrics. For example, IRAP measures at the level of the single, overall IRAP Effect score, list-level Trial Type score, and D-IRAP scores down to the level of individual stimuli were utilized for purposes of analysis. For data entry task performance, metrics including total number of sessions completed, average correct trials per session, overall percentage of correct responding, number of sessions completed following failure to meet the minimum work requirement, and total amount of revenue earned were utilized for purposes of analysis. Prior to the conduct of statistical analyses to determine the predictive validity of IRAP results with respect to data entry task performance, visual inspection of the data was employed in order to determine the extent to which it appeared that IRAP results bore some relation with performance results.

With respect to IRAP results, it was observed that one participant exhibited a Pro-hard work bias (PP2-3, MT-IRAP), while other participants exhibited a Pro-easy work bias (PP2-1, MT-IRAP, and PP2-4, traditional IRAP). Two participants, PP2-2 and PP2-5, exhibited somewhat mixed results on the traditional IRAP. Similarly, some participants exhibited greater performance on the data entry task, in terms of the various metrics described above, while some participants exhibited lesser performance on the task. The following section describes how these measures relate for each participant, based on visual inspection of the data.

For participant PP2-1, who exhibited a Pro-Easy work bias in the MT-IRAP, it was observed that performance on the data entry task was very brief, relative to other participants, having only completed four 2-min sessions before choosing to quit the task. This participant averaged 9.75 correct trials per 2-min session, which was one of the lowest rates among all participants, and did not persist at the task long enough to have encountered a minimum work requirement which was not able to be met. In addition, upon drilling down the IRAP measures to the level of individual stimuli, this participant exhibited a roughly neutral D-IRAP score for the Like→Persistence trial type (0.13) and a very strong Anti-Persistence bias (-1.77) for the Dislike→Persistence trial type (-0.82 for Persistence overall when averaged across both trial types). In addition, the D-IRAP score for Like→Quitting indicated a strong Anti-Quitting bias (1.11), while for Dislike→Quitting the D-IRAP score was -0.49, which indicated a moderate Pro-Quitting bias. A weak-to-moderate Anti-Quitting bias was obtained when the two were averaged to yield a single D-IRAP score for Quitting. In sum, this participant earned $1.17 throughout the data entry task, which was one of the lowest amounts earned among all participants. Based on these descriptive analyses, it appears that in the case of this participant, Pro-Easy work biases, as measured by the MT-IRAP, were at least correlated with, if not predictive of, quitting the data entry task quickly (i.e., not persisting at the task), performing at a low rate, in terms of average correct trials per session, and earning a low amount of revenue during the work task. The D-IRAP score for the individual stimulus of Persistence also appeared to correlate well with these work performance measures and the duration of persistence at the work task.

Participant PP2-4 also exhibited largely Pro-Easy work biases, as measured through the traditional IRAP assessment, at the levels of both lists of stimuli (i.e., Trial Types) and the overall IRAP Effect (i.e., averaged Trial Types). This participant's work performance measures consisted of a large number of 2-min sessions completed (19), however, at no point did this participant ever meet the minimum work requirement for a given 2-min session. Thus, while this participant did persist at the task and complete a large number of 2-min sessions following failure to meet the minimum requirement (18), this participant's performance, in terms of average correct trials per 2-min session (4.1) and total revenue generated during the task ($0.00) were the lowest among all participants. Interestingly, when this participant's IRAP results were examined at the level of individual stimuli, it was observed that while there was a general Pro-Easy work bias across most of the "hard work" stimuli, this participant exhibited a strong Pro-Persistence bias for Like→Persistence (0.75) and Dislike→Persistence (1.67) trial types, which averaged out to a strong Pro-Persistence bias for Persistence overall (1.21). The individual stimulus D-IRAP scores for Quitting were more ambiguous, with a strong Pro-Quitting bias for Like→Quitting (-1.32) and a moderate Anti-Quitting bias for Dislike→Quitting (0.67), which yielded a weak-to-moderate Pro-Quitting bias for the stimulus of Quitting overall (-0.32). Therefore, for this participant in particular, the Trial Type and overall IRAP Effect scores appeared to correlate well with performance, in terms of average correct trials and revenue earned, both of which were lowest among all participants. And although this participant persisted at the work task for some time, even after having failed to meet the minimum work criterion in every session, this correlated well the IRAP scores for the individual stimulus of Persistence, as was also observed for participant PP2-1, as noted above.

Participant PP2-3 demonstrated a general Pro-Hard work bias, as measured by the MT-IRAP, at the levels of lists of stimuli and Trial Type, as well as overall IRAP Effect. This participant's work performance consisted of completing 23 2-min work sessions, with an average of 14.6 correct trials per 2-min session, both of which were in the upper range among all participants. This participant did contact failure to meet the minimum work requirement and completed an additional three sessions following that failure. Lastly, this participant earned $8.22 during the work task, which was the second-highest among all participants. Upon examining the IRAP scores at the level of individual stimulus, this participant exhibited a very strong Pro-Persistence bias for the Like→Persistence trial type (2.23) and a weak Pro-Persistence bias for the Dislike→Persistence trial type (0.28). When averaged together, these scores yielded a strong Pro-Persistence bias (1.13) for the Persistence stimulus overall.

Additionally, when looking at individual "easy work" stimuli, although there were some Pro-Easy work biases detected among some of the specific stimuli (i.e., Easy, Leisurely, Relaxing, and Calm), as well as some neutral or absence of bias for some stimuli (i.e., Slow and Simple), there was an Anti-Quitting bias present for the stimulus of Quitting, based on a weak-to-moderate Anti-Quitting bias for the Like→Quitting trial type (0.31) and a strong Anti-Quitting bias (1.06) for the Dislike→Quitting trial type, with a moderate Anti-Quitting bias when the two were averaged (0.69). Thus, visual inspection of these descriptive data suggests that the IRAP measures for the overall IRAP Effect, the list-level Trial Type effects (specifically, Trial-Types 1 and 4), and the individual stimuli correlated well with the work performance measures, specifically those of sessions completed, average correct responses per 2-min session, persistence of the task after contacting failure to meet minimum work requirements, and revenue earned during the task.

Participant PP2-2 was one of the two participants who demonstrated largely mixed or inconclusive results on the IRAP measure (traditional IRAP). Specifically, a Pro-Hard work bias was exhibited for the two Trial Types involving "hard work" words (i.e., Trial Types 1 and 3), while a Pro-Easy work bias was seen for the other two Trial-Types, which involved "easy work" words (i.e., Trial Types 2 and 4). When averaged across each other to yield the overall IRAP Effect score, it was roughly neutral with a slight Pro-Easy work bias (-0.13). Regarding the work task, this participant completed 25 work sessions, which was the most completed by any work participant, with an average of 16.3 correct trials per 2-min session, which was highest among participants. This participant earned $11.76 in revenue throughout the task, which was the greatest among all participants, and did not fail to meet the minimum work requirement at any time, prior to choosing to quit the work task. The two individual stimuli from the IRAP assessment which were most directly related to the work task (i.e., Persistence and Quitting) yielded either contradictory, mixed results, as in the case of Persistence (Like→Persistence = 0.46, Dislike→Persistence = -0.57, combined Persistence = -0.06), or a strong, Pro-Quitting work bias, as in the case of Quitting (Like→Quitting = -1.28, Dislike→Quitting = -0.39, combined Quitting = -0.84). In either case, it did not appear that the individual stimuli D-IRAP scores for Persistence or Quitting, as measured by the traditional IRAP, correlated with or predicted the participant's performance on the data entry work task. Given the largely mixed results of the IRAP at the Trial Type list level of analysis, it cannot be said that those results correlated with or predicted work performance.

Participant PP2-5 was the other participant who exhibited largely mixed, inconclusive results on the IRAP, which again was the traditional IRAP. Similar to participant PP2-2, this participant demonstrated a Pro-Hard work bias on the two Trial Types involving "hard work" stimuli (Trial Types 1 and 3), and a Pro-Easy work bias on those Trial Types involving "easy work" stimuli (Trial Types 2 and 4). When averaged together to produce a single IRAP Effect score, the disparate Trial Type D-IRAP scores cancelled each other out and yielded a neutral IRAP Effect score (0.04). With respect to the work task, this participant's performance was among the lower ranges of the various measures. In particular, this participant completed 11 2-min sessions, which was the second fewest among all participants. This participant completed an average of 10.7 correct trials per session, which was the median among participants, with an overall average of 86% accuracy throughout the entire task, which was second-lowest among participants. This participant failed to meet the minimum work requirement during the eighth session and completed another three sessions following that failure.

Over the course of the entire task, this participant earned $2.19 in revenue, which was lower range among all participants. In terms of the D-IRAP scores of individual stimuli, this participant's scores for Persistence indicated a strong Pro-Persistence bias (Like→Persistence = 6.01, Dislike→Persistence = 1.41, combined Persistence = 3.71), while the D-IRAP scores for Quitting were indicative of a Pro-Quitting bias (Like→Quitting = -0.23, Dislike→Quitting = -1.15, combined Quitting = -0.69). It is difficult to estimate through descriptive statistics alone the extent to which any IRAP results correlate with work performance for this participant. Although this participant did persist at the task for several sessions after first failing to meet the minimum work requirement, which correlates with the strong Pro-Persistence bias seen with the Persistence stimulus, this participant's overall work performance was among the poorest of the participants, in terms of numbers of sessions completed, average correct trials per session, overall work task accuracy, and revenue earned.

Discussion

The purpose of Experiment II was to determine the extent to which the results of either of the IRAP instruments correlated with or predicted other patterns of behavior pertaining to a simulated data entry work task. While there are several ways to analyze

IRAP results (e.g., overall IRAP Effect, Trial Type list level effects, and results for individual stimuli) and a number of pertinent measures associated with the work task, providing for a multitude of different ways to attempt to draw correlation and prediction between the two, thus far it appears that for a majority of participants (i.e., three of five) there was some amount of correlation between IRAP results for a given participant and that participant's performance on the work task. Specifically, in terms of the two groups of participants which were exposed to only IRAP or the other, the two participants in the MT-IRAP group exhibited clear IRAP results which appeared to correlate well with their respective work task performance. Of the three participants in the traditional IRAP group, one participant produced consistent results, in terms of implicit bias, which correlated well with work task performance, while the other two participants produced evidently inconsistent results, resulting in no apparent correlation with work task performance. Further statistical analyses utilizing statistical models to establish incremental predictive validity will help to further elucidate the relationship between the IRAP results pertaining to attitudes towards work and overt patterns of performance in the analog work setting.

Although the preliminary results of Experiment II indicate that the MT-IRAP produces results which correlate better with other overt patterns of behavior in an analog work setting, the results of Experiment I, which was conducted concurrently with Experiment II, emphasized the need to gather more data (i.e., from both IRAP iterations) for each participant, in order to improve analyses concerning predictive utility. It should be noted that the two participants whose IRAP data were mixed and somewhat inconclusive both completed only the traditional IRAP assessment. Given the results of

Experiment I of the present study, in which the traditional IRAP appeared to produce less consistent results relative to the MT-IRAP, it remains unclear the extent to which the MT-IRAP would have produced results which correlated better with the work task performance of those participants. In light of the outcomes of both Experiments I and II, it was decided that participants in Experiment III would complete both IRAP instruments, rather than only one, making Experiment III more of a within-subject comparison than was the case with Experiment II.

Thus, the purpose of Experiment III was to further investigate the extent to which either IRAP instrument offered any prediction of how participants would behave, in terms of other overt patterns of behavior, in the analog work setting. In order to do so, Experiment III was conducted on a within-subject basis, in which each participant completed both IRAP assessments as well as the simulated work task, in order to determine which IRAP yielded greater predictive validity using the same participant sample.

Experiment III

Method

Participants, Setting, and Apparatus

Participants in Experiment III were 22 undergraduate students at the University of Nevada, Reno, who did not participate in Experiments I or II. Participants signed up through the Sona-system online subject pool management system, in which participants were instructed they must have at least two years cumulative work experience in either part-time or full-time job positions, in order to participate. The study took place in the same small laboratory rooms on the university campus in which

Experiments I and II took place. Similarly, the same set of PC desktop computers from Experiments I and II were utilized to conduct Experiment III.

Independent Variables

The independent variables for Experiment III were the two IRAP instruments to which participants were exposed and the increasing work demands placed on participants during the simulated data entry work task. More specifically, during the simulated work task participants had to perform at an increasingly demanding minimum level of performance, in terms of correct trials per 2-min session. Periodically, the work demands per session increased.

Dependent Measures

The primary dependent measures were the response latencies as measured by the IRAP instruments (as in Experiments I and II) and the performance measures associated with the work task. Specifically, primary measures for the work task consisted of number of 2-min work sessions completed; average number of correct trials per 2-min session; and persistence, in terms of number of sessions completed during the work task. Secondary dependent measures consisted of Likert-type self-report ratings of how stressful and demanding the work task was at different points in time throughout the task.

Research Design and Procedure

The research design was a between-groups analysis of the extent to which IRAP scores from the different versions of the IRAP correlated with and predicted the dependent measures from the simulated work task; however the between-groups data were drawn from the same sample of participants. More specifically, each participant was exposed to both the traditional IRAP (A) and MT-IRAP (B), as well as the work task

(C), and their results from both IRAPs and the work task were analyzed in separate correlational analyses to determine the extent to which either set of IRAP results correlated with the work task results. Certain aspects of the sequence of exposures were counterbalanced across participants, in order to control for any potential sequence effects (see Figure 12). All of the participants completed the IRAP assessments first, followed by the work task. Half of participants, completed the traditional IRAP first (ABC sequence), while the other half completed the MT-IRAP first (BAC sequence). The counterbalancing of the IRAP assessments was done as it was in Experiment I, in order to identify and control for possible sequence effects with respect to completing two IRAP assessments consecutively (even though there was no discernible sequence effect observed in Experiment I).

The two IRAP assessment instruments and the simulated data entry work task were identical in all procedural details as they were in Experiments I and II. As indicated above, the only difference between Experiments II and III was that all participants in Experiment III were exposed to both IRAP instruments, instead of only one, in addition to the data entry work task. Otherwise, see Experiments I and II above for specific details regarding the experimental procedure for Experiment III.

Results

The results for Experiment III are presented below, beginning with the direct comparison of the convergence of the two IRAP assessment results, as was done for Experiment I, which thereby served as a replication of Experiment I.

List-Level Analyses

The first comparisons of convergence among the two IRAPs were conducted at the level of lists of stimuli (i.e., hard work and easy work words). Table 6 presents the Pearson correlation coefficients *r* for each participant, correlated across each of the four Trial Types. It should be noted that each correlation coefficient listed for each participant is based only on four observations for each correlation (n = 4), given that there are only four Trial Type measures per participant. As such, these correlation values must be interpreted cautiously, since there are so few data in each analysis. Inspection of these correlations indicated that of 22 participants, six participants exhibited correlations which were moderate-to-strong, positive correlations (i.e., *r* > .50; P5, P20, P21, P24, P25, and P29). Seven participants (P1, P2, P3, P12, P17, P19, and P30) exhibited strong, negative correlations (i.e., *r* < -.50). The remainder of individual participants' IRAP scores resulted in correlation coefficients between -.50 and .50, of which six participants (P4, P8, P10, P11, P13, and P22) were positive correlations and three (P28, P30, and P32) were negative.

Taken together, the correlation coefficient analyses described above indicated that 12 of 22 participants exhibited positive correlations of varying strengths between their two IRAP assessment results and 10 participants exhibited negative correlations of varying strengths. As mentioned previously, these individual participant correlations must be interpreted cautiously, therefore, a single correlation coefficient was calculated across all participants' scores, in which each Trial Type was compared for all participants. In other words, Trial Types 1 – 4 for each IRAP for all participants were entered into the correlation analysis, which produced n = 88 observations in each data set for each IRAP, allowing for a more robust correlation coefficient analysis, in which more confidence could be placed. The result was a correlation coefficient of r = -.01, or no correlation between the two IRAP instruments. Based on the coefficient value (-.01), a test of inferential statistical significance was not conducted for this correlation, given that it indicated no correlation. Broadly, based on both the individual participant and group analyses described above, the two IRAP assessment instruments did not produce convergent results across the same participant pool.

As was noted in Experiment I, the difference in error rates observed across the four different Trial Types was substantial, with the third and fourth Trial Types (i.e., those containing “Dislike” as the evaluative stimulus) having a greater frequency of errors. Results in Experiment III demonstrated a similar pattern, thus the frequency of errors in the traditional IRAP was statistically analyzed to evaluate whether there were significantly more errors among any of the different Trial Types. The mean errors and associated variances were calculated for each Trial Type in the traditional IRAP and the following values were obtained: Trial Type 1, M = 3.46 errors, s^2 = 7.65; Trial Type 2, M = 3.88, s^2 = 5.94; Trial Type 3, M = 7.29, s^2 = 20.91; Trial Type 4, M = 8.21, s^2 = 38.09. These values were subjected to a single-factor ANOVA, which yielded an omnibus result of F(3, 92) = 7.57, p = .0001. Follow up comparisons indicated that average errors between Trial Types 1 and 2 and average errors between Trial Types 3 and 4 were not significantly different from each other. However, average errors for Trial Type 1 were significantly lower than average errors for both Trial Type 3 and Trial Type 4, and the same was found to be true for Trial Type 2. In other words, the Trial Types in which “Like” was the evaluative stimulus had, on average, significantly fewer errors across the group of the participants than did the Trial Types which contained “Dislike” as the evaluative stimulus.

A similar analysis to determine differences in frequency of errors was conducted with the MT-IRAP data. Prior to the analysis, problematic data were removed from the data set. The source of these problematic data was the fact that participants could potentially give different answers to particular trials at different times during the MT-IRAP (e.g., first versus last test block). This could be problematic, given that the first test block of the MT-IRAP was used to determine how each participant responded to each individual combination of stimuli, as opposed to forcing participants to respond in a certain manner, as the traditional IRAP did. The MT-IRAP operated such that when a participant provided certain responses to particular trials during the first test block—after which those responses were recorded as the “correct” responses for that participant—providing different responses to the same trials in later blocks was recognized as an error. The drawback to this procedure occurred when a participant gave certain responses to certain trials in the first test block, but then gave different answers to those same trials in later test blocks. There were two circumstances under which participants’ responses varied from the first test block to later blocks, namely, when a participant simply “changed his or her mind” about how to respond to particular trials (e.g., Like→Demanding→Truth), or when a participant consistently answered incorrectly during the first test block (i.e., making mistakes) while the “correct” answers were being established, such that the program recorded the “incorrect” response for that participant as the “correct” response. In such an instance, when the “correct” response that the participant “meant” to give was emitted in later test blocks, the program recognized these as errors and was not able to identify the shift in responding and identify the new "correct" response as such. When either case of providing different answers at different times occurred, the MT-IRAP program responded by presenting the same trial (e.g., Like→Demanding→Truth) repeatedly at the end of the test block (after other trials had been answered "correctly"), until the response it had initially recorded as "correct" was finally emitted by the participant. This procedural difficulty occurred at varying frequencies for different participants, whereby some participants did not experience this at all, and others experienced this for several target stimuli.

In order to account for the potential errors described above, the data for any stimulus for which this pattern of responding and errors was observed for a given participant were removed from analyses. Data for particular stimuli were removed on the basis of repeated trials of the same exact trial (e.g., Like→Demanding→Truth) appearing in consecutive presentations at the end of a test block with consecutive "incorrect" responses having been repeatedly emitted, until the "correct" response was finally emitted to end the test block. For the purposes of list-level analyses among the two instruments, all data for a given stimulus were removed on this basis, even if only one Trial Type of the stimulus (e.g., Dislike→Demanding) exhibited a high frequency of errors, while the other Trial Type (e.g., Like→Demanding) did not.

Once problematic error data were identified and removed, as described above, an analysis of differential error rates across the four different Trial Types was conducted for MT-IRAP data. Descriptive statistics indicated that the mean number of errors for Trial Type 1 was M = 4.57, s^2 = 32.26; Trial Type 2 M = 4.90, s^2 = 32.09; Trial Type 3 M = 10.71, s^2 = 115.71; and Trial Type 4 M = 10.52, s^2 = 65.66. These values were subjected to a single-factor ANOVA to test for statistical significance, which yielded *F*(3, 80) = 3.95, *p* = .01. Follow up post hoc comparisons indicated that frequency of errors for Trial Types 1 and 2 were not significantly different from each other, but that both Trial Types 1 and 2 were each significantly different from Trial Types 3 and 4. Trial Types 3 and 4, however, were not significantly different from each other. These results replicated the similar pattern of errors observed in the traditional IRAP, in which the Trial Types containing the evaluative stimulus "Like" had significantly fewer errors than the Trial Types which contained the evaluative stimulus "Dislike".

Given that the above analyses of errors indicated that Trial Types 3 and 4 were prone to significantly more error responses from participants, a subsequent correlational analysis between the two IRAP instruments was conducted using only Trial Types 1 and 2, i.e., only those trials containing "Like" as the evaluative stimulus. The result was a coefficient of *r* = .12, which was improved over the -.01 observed with all four Trial Types included, however, it failed to reach statistical significance, at *p* > .20.

In summary, list-level analyses of the two IRAPs indicated a wide range of correlation coefficients at the level of individual participants, which suggested an absence of any reliable convergence among the two instruments. When all participants' data were aggregated into a single correlational analysis, this too indicated no correlation or convergence among the two IRAP instruments. It was observed, however, that there were significantly different error rates for participants on "Dislike" trials in the IRAPs, and when only "Like" trials were analyzed for convergence, the correlation coefficient between the two IRAP instruments improved to *r* = .12, however, it was not a statistically significant correlation between the two instruments.

Individual Stimulus Analyses

Beyond comparing the two IRAP instruments at the level of lists of stimuli, results of the two assessment tools were also analyzed for convergence and correlation at the level of the individual stimuli. These analyses were conducted at both the individual participant level, as well as at the group level. The within-participant (i.e., individual participant) correlational analyses could be interpreted with more confidence at the level of individual stimuli, since there were 16 individual stimuli tested in each IRAP, such that there were n = 32 data (16 target stimuli each paired with the two evaluative stimuli) entered into the Pearson correlation coefficient calculation (as opposed to only n = 4 data in the calculations for entire lists of stimuli, as described in the list-level analyses above). Table 7 displays the Pearson's product-moment correlation coefficient *r* between the two IRAP results for each participant, analyzed at the level of individual stimuli. Figures 13 through 15 display representative graphic depictions of the IRAP scores for each stimulus in each of the IRAP instruments for a representative sample of participants (P20, P3, and P4). At this level of analysis, no correlation coefficients for any of the participants exceeded .50 in either direction (positive or negative), indicating that all correlations were weaker than were observed at the level of lists of stimuli. However, since there was a much larger data set entered into each correlational analysis, two of the correlation coefficients represented statistically significant correlations. Specifically, Participant 20's scores produced *r* = .49, *p* = .005 if analyzed as a directional correlational analysis (similar to a one-tailed test of significance), and *p* = .01 when analyzed non-directionally (similar to a two-tailed test). Additionally, Participant 3's scores produced *r* = -.38, with a directional *p* = .02, and non-directional *p* = .05.

Lastly, IRAP scores for Participant 8 produced a correlation coefficient of *r* = .31, which achieved statistical significance in a directional analysis at *p* = .04, but missed statistical significance in a non-directional analysis at *p* = .08. Overall, five participants demonstrated correlation coefficients of greater than .25, indicating weak-to-moderate positive correlations, but only those noted above reached statistical significance. Two participants produced correlation coefficients of less than -.25, indicating weak-to-moderate negative correlations, with only one being statistically significant, and all other participants' correlations were between -.25 and .25. Of all 22 participants' correlations, 15 were positive to at least some extent, and seven were negative to at least some extent.

A correlational analysis was also conducted at the level of individual stimuli across the entire group of participants, whereby all of the individual stimulus IRAP scores for all of the participants were entered into a single correlational analysis, which produced a result of *r* = .05. The correlation coefficient did not achieve statistical significance, with a directional *p* value of .08. Given the aforementioned problems associated with the "Dislike" trial types, in terms of frequency of errors, a similar correlational analysis was conducted using only the "Like" trial IRAP scores, which produced a coefficient of *r* = .10. This analysis achieved statistical significance when analyzed directionally, at *p* = .04, however, when analyzed non-directionally, it only approached significance, at *p* = .07.

Upon visual inspection of the IRAP scores of individual stimuli within individual participants, it became apparent that the stimuli chosen to represent "hard work" and "easy work", which were chosen based upon topographical and semantic similarity, were not functionally equivalent, as measured by the IRAP instruments. Common examples of the “hard work” words included different attitudes, as measured by the IRAPs, to words such as Effortful and Stressful, in which pro-Effortful and anti-Stressful biases were demonstrated. Similarly, the stimulus of Persistence often had a pro-Persistence valence associated with it, even when many of the other “hard work” words were found to have anti-biases associated with them. On the “easy work” side of the stimuli, it was often found that the stimulus “Lazy” had an anti-Lazy bias exhibited by participants, as did the stimulus of “Quitting”, whereas “Relaxing” often had a pro-Relaxing attitude associated with it, as did the stimulus of “Leisurely”.

Explicit Measures

In addition to completing each of the two implicit IRAP measures, participants also completed an explicit survey measure at the beginning of the experimental session, prior to completing the IRAPs. The explicit survey was a Likert-type survey measure (e.g., Strongly Agree, Somewhat Agree, Strongly Disagree, etc.), which asked participants to provide their explicit attitudes to the concepts proposed in each question. The questions were developed by the researchers to provide an explicit counterpart to the implicit attitudes presented in the IRAP assessments (see Appendix B). As such, each question incorporated one of the target stimuli from the IRAP assessments. Explicit scores were therefore obtained at the level of individual stimuli (i.e., each survey question), as well as at the overall level (similar to the list level IRAP scores), in which all scores for all questions were averaged into a single explicit survey score.

Each survey question was scored with a positive or negative value, with a maximum range of -3 to 3 corresponding to the Likert-type answer options (0 was not an option), such that a positive score indicated a pro-Hard work bias, and a negative score indicated a pro-Easy work bias, as was the case with the IRAP scores. Table 9 presents the average scores for each question, with the corresponding target stimulus for each listed in parentheses. On average, each survey question and corresponding target stimulus demonstrated a pro-Hard work explicit bias (i.e., score greater than zero) among the group of participants, with exception of only two questions: questions 10 and 11, pertaining to the target concepts of “Easy” and “Calm”, respectively. In addition, all participants demonstrated a pro-Hard work explicit bias on the survey, when the scores for each question were averaged for each participant. In summary, participants expressed varying degrees of pro-Hard work biases through the explicit attitude assessment, and few participants expressed pro-Easy work explicit attitudes on average, with some exceptions to a few question items.

The explicit survey scores were correlated with the implicit scores from each of the IRAPs, in order to evaluate any convergence among the explicit and implicit measures. Table 10 presents the correlation coefficients between the explicit survey measures and the IRAP assessments scores from each IRAP, at both the stimulus and list levels. The correlational analyses revealed a wide range of correlation coefficient, both positive and negative, the majority of which were fairly weak and statistically non-significant. There were three correlation coefficients between explicit and implicit measures which did achieve statistical significance; those were Difficult MT-IRAP, r = -.46, p < .05; Demanding MT-IRAP, r = -.50, p < .05; Easy-2 MT-IRAP, r = .46, p < .05; all p-values refer to non-directional, two-way analyses. There were a couple explicit/implicit correlations which approached significance, in particular, Effort TR-IRAP, r = .43, p = .07; and Lazy MT-IRAP, r = .41, p = .08. The overall explicit survey scores were subjected to correlational analyses with the list-level IRAP scores for Trial Types 1 and 2 (i.e., Like-Hard work and Like-Easy work, respectively), however, none of the correlation coefficients were statistically significant.

Work Task Measures

The dependent measures collected during the analog data entry work task are displayed in Table 8. The primary dependent measure within this data set was that of number of 2-min work sessions completed by each participant. Additional measures included the average number of correct trials per 2-min work session, the total percentage of correct trials across all 2-min sessions, the number of 2-min sessions completed after failure to meet the minimum work criterion, revenue earned, and self-reported levels of stress and task demand during the work task. The number of sessions completed ranged from three to 36, with a mean of *M* = 13.35, *SD* = 9.00, and median = 12. Given that the largest value of 36 was much greater than any of the other values in the data set and was greater than 2.5 standard deviations above the mean, a test was conducted to determine if it was a statistical outlier.

Given that certain measures of central tendency, such as mean and standard deviation, are sensitive to influence by outliers, a more robust measure using the median of the dataset was utilized, which is appropriate when few or only one outlier is suspected. The median absolute deviation (MAD) procedure was used to determine if any outliers were present in the data set. A conservative test statistic criterion of 3.0 was used as cutoff to determine if any values represented outliers. All values in the data set produced test statistics of < 2.5, with the exception of the 36 datum, which produced a test statistic of 3.43, and was therefore treated as an outlier and removed from the analyses described below.

Predictive Validity Analyses

Bivariate linear regression evaluates the prediction of a "dependent" variable from an "independent" variable or, in the case of multiple regression, set of "independent" variables, although the predictor variables need not be independent variables in the formal sense (Bryant, 1960; Salkind & Green, 2011). Single and multiple regression analyses were conducted using various IRAP scores from the two IRAP instruments as predictors of different work task measures, including number of sessions completed (i.e., behavioral persistence at the task) and average correct trials completed. Number of sessions completed was envisioned as the primary dependent measure from the work task to be predicted by IRAP scores, given that two of the stimuli assessed by the IRAP ("Persistence" and "Quitting") were assumed to be functionally related to the behavioral persistence measured by the analog work task. Since each participant had the option to quit the work task at any time during the work task, it was assumed that actual quitting and persisting at the task would be related to IRAP scores pertaining to those word stimuli to at least some extent. Also, more broadly, it was assumed that measured implicit attitudes toward the entire lists of "hard work" and "easy work" stimuli might be predictive of some more global measure of work task performance, such as revenue earned or average correct trials completed.

Beginning with the IRAP scores at the level of lists of stimuli, none of the four Trial Types from either of the two IRAP instruments correlated with or predicted number of sessions completed by participants. When the list-level IRAP scores for each Trial

Type were correlated with the average correct trials per session during the work task, as a broad measure of task performance, it was found that a weak positive correlation existed for the Like-Hard Work trials (Trial Type 1) of the MT-IRAP and average correct trials. As seen in Figures 16 and 17, this weak-to-moderate correlation was present with the MT-IRAP scores, however, not with the traditional IRAP scores. Specifically, a correlation coefficient of $r = .32$ was obtained, with an $R^2 = .11$ for the MT-IRAP. While this indicated some level of prediction by the MT-IRAP scores for Trial Type 1, when predicting average correct trials during the work task, a test for statistical significance indicated that it was not a statistically significant correlation ($p = .07$ directional, $p = .14$ non-directional).

Moving to the level of individual stimulus IRAP scores, the scores of Persistence and Quitting were examined to evaluate the extent to which they predicted number of work sessions completed by participants. In particular, given the difficulty observed in responding to IRAP trials containing the evaluative stimulus “Dislike”, it was predicted that the Like→Persistence trials would yield the greatest predictive validity of any IRAP scores. As seen in Figure 18, when the Like→Persistence IRAP scores for the traditional IRAP were correlated with number of work sessions completed, it yielded a coefficient of $r = .24$, $R^2 = .06$, which when subjected to null hypothesis testing that there is no correlation in the population, failed to achieve significance, given a directional $p = .14$ and non-directional $p = .28$. When the same IRAP trials of Like→Persistence from the MT-IRAP were subjected to a correlational analysis with number of work sessions completed (Figure 19), an $r = .56$ was obtained, $R^2 = .32$, and a directional $p = .003$ and non-directional p = .006 indicated that the null hypothesis of no significant correlation inferred in the total population was rejected.

An additional inferential statistical analysis was conducted in order to determine whether the two correlation coefficients noted above were statistically significantly different from each other. This was done using a Fisher *r*-to-*z* transformation, which allowed for a comparison of the two correlation coefficients, though this is typically done with two independent samples. The analysis indicated that the two correlation coefficients were not significantly different from one another, although the MT-IRAP analysis alone was clearly significantly different from the null hypothesis that it did not predict number of work sessions completed. A "what-if" analysis was conducted, whereby it was determined that if the correlation coefficients were to remain approximately the same as indicated in the above analysis, then approximately n = 37 participants would be required in order for the two correlation coefficients to reach a statistically significant difference from one another.

Given the requirement needed to achieve statistical significance when the two IRAP correlations are directly compared, as described above, a post hoc power analysis was conducted on the MT-IRAP correlational analysis in order to determine what level of power had been achieved, given the data set. Given the single predictor of Like→Persistence IRAP scores, an observed R^2 = .32, probability level of .05, and a sample size of n = 22, it was determined that a power of .87 had been achieved for that particular analysis. An *a priori* calculation of power for correlational analyses indicated that in order to achieve .90 power, given the observed r and R^2 values of the MT-IRAP scores in the present study, a sample size of 25 participants is required (just three more than the 22 included in the current analysis).

A number of multiple regression analyses (for example, using both “Like” and “Dislike” Trial Types for “Persistence” or “Quitting”, combining different Trial Types of “Persistence” and “Quitting”, or combining “Persistence” or “Quitting” scores with list-level IRAP scores) were conducted using various IRAP scores from both instruments to predict number of work sessions completed or other work task measures, however, although some additional predictors increased the R^2 variance accounted for in the predicted variable, none of the additional predictors added a statistically significant increase in R^2 variance accounted for. Therefore, no multiple regression analyses are reported.

When the analyses conducted for Like→Persistence IRAP trials and work task persistence were parsed based on sequence of IRAP exposures for participants, a sequence effect was observed. In particular, it was noted that whichever of the two instruments was administered first yielded greater predictive validity than the instrument which followed it, as shown in Figures 20-23. More specifically, for the group of participants who completed the traditional IRAP first, it was more predictive of work sessions completed (r = .51, R^2 = .27, one-way p = .07) than was the MT-IRAP (r = .44, R^2 = .19, one-way p = .10), though neither was a statistically significant correlation. Similarly, for the group of participants who completed the MT-IRAP first, it was more predictive of work sessions completed (r = .75, R^2 = .56, one-way p = .002, two-way p = .005) and was highly significantly predictive, whereas the traditional IRAP was not at all predictive of work sessions completed (r = .02, R^2 = .00).

A similar investigation of the predictive validity analyses which were conducted for the Like-Hard work list IRAP scores and average correct trials from the work task did not reveal the similar sequence effect. In particular, even when the traditional IRAP was administered first (ABC sequence), it was not at all predictive of average correct trials performance (r = -.07), whereas the MT-IRAP was predictive of participants' performance (r = .43, R^2 = .19, p > .10) to some extent. In instances that the MT-IRAP was administered first, it was more predictive (r = .20, R^2 = .04) than was the traditional IRAP (r = -.11), but not as predictive as when the MT-IRAP was administered following the traditional IRAP, as described above.

Predictive validity analyses were also conducted utilizing the explicit survey scores to predict the work task measures, as was done with implicit IRAP scores. Figures 24 and 25 graphically depict these analyses, the first of which was similar to the list-level IRAP scores used to predict general work task performance, measured as average correct trials per 2-min session (see Figure 17), however, in this case the overall explicit survey scores were instead used to predict average correct trials per 2-min session, as shown in Figure 24. A correlation coefficient of r = .42 was found, which approached statistical significance in a two-way analysis with p = .07 (p = .04 one-way), R^2 = .17. Thus, the average score from the entire explicit survey measure moderately predicted participants' performance on the work task, in terms of average correct trials completed. The second analysis conducted incorporated the explicit survey scores from the question pertaining to Persistence and the number-of-sessions measure collected from the work task. The correlation coefficient was r = .05, with an R^2 = .00. Thus, there was no prediction of persistence at the work task, measured in terms of number of sessions completed, by the explicit measures for the Persistence item

In summary, there were only a couple of regression analyses that indicated substantial predictive capability of IRAP scores with respect to any of the work task measures. Ultimately, these consisted of single regression, or correlation, analyses, since no multiple regression analyses with multiple predictors indicated multiple statistically significant combinations of predictors. Of the single regressions that indicated some amount of predictive validity, the two strongest were observed with MT-IRAP scores, specifically, the MT-IRAP scores for the Like→Persistence trials strongly predicted persistence in the work task (i.e., number of sessions completed), and the MT-IRAP scores for the Like→Hard Work words list moderately predicted performance in the work task, measured as average correct trials per session. The traditional IRAP scores for Like→Persistence trials exhibited a weak prediction of persistence at the work task, and this was the only prediction observed for any traditional IRAP scores. Additionally, a sequence effect was observed for one of the correlation analyses (Like→Persistence individual stimulus analysis), but not the other (Like→Hard Work list).

Discussion

The purpose of Experiment III was twofold. Firstly, it served as a replication of Experiment I, whereby both IRAP assessment instruments were evaluated to determine to the extent to which they produced convergent, correlated results on a within-subject basis, as well as across the entire group of those same participants. Secondly, Experiment III sought to determine which IRAP instrument, if any, would yield any predictive validity, in terms of predicting overtly observable behaviors of interest as measured within an analog organizational work task. With respect to the first primary research question, Experiment III did replicate Experiment I, insofar as there was a wide range of correlations observed between the two IRAP instruments, using the same exact sets of stimuli, on a within-subject basis. The correlation coefficients ranged from strong negative to strong positive correlations, with many values falling somewhere between, including many participants for whom there was no correlation between the two instruments. When the data for all participants were combined into a single correlational analysis, no correlation between the two instruments was observed, and this was found to be the case in both Experiments I and III.

The second research question investigated in Experiment III, and perhaps the more important of the two, was whether either IRAP instrument was able to predict other objectively observable behavioral patterns of interest; in this case, various metrics pertaining to the performance of participants during an analog organizational work task. Of the various measures obtained from the IRAP assessments, including D-IRAP scores for combined lists of stimuli as well as for individual stimuli, most did not yield any predictive validity with respect to the number of different measures obtained from the analog work task. In fact, the only IRAP results which yielded a strong effect as far as predicting work task measures were the D-IRAP scores associated with the specific combination of stimuli “Like” and “Persistence”. These IRAP scores predicted the behavioral work task measure considered as “persistence”, which consisted of the measure of how many 2-min work sessions were completed by each participant during the simulated work task. Importantly, only the IRAP scores for the MT-IRAP were able to provide significant prediction of persistence at the work task, whereas the traditional

IRAP scores for the same stimuli produced only a small, non-significant prediction of the behavior in question.

The interpretation of these findings is complicated by a number of factors which are presumed to affect the extent to which a latency-based measurement tool, such as the IRAP (and others like it, such as the IAT), can provide a valid set of data for analytical purposes. Many of such factors appear, based on the empirical observations made within the present study, to be related primarily to procedural aspects of the assessment itself. For example, as noted earlier, a key difference in the procedures of the two IRAP assessments is the handling of incorrect or error responses and their associated latencies. The traditional IRAP presents each combination of evaluative and target stimuli just once in each block of trials, and if an incorrect response is emitted on a given trial, then the latency of that incorrect response is kept and used in the data set to be analyzed. Typically, this is handled in a manner whereby any participant whose percentage of incorrect responses falls below 75% for any given test block, the data for that test block (and its corresponding paired test block) are removed from the subsequent analyses. In the event that a participant's percentage of correct responding falls below 75% for any two blocks, that participant's data are removed entirely from all analyses. The traditional IRAP data from Experiment III empirically demonstrated the potential problems with this procedure for handling error responses and associated error latencies, in that the majority of participants' percentage of correct responding never fell below 75% correct for a given block, however, responding for certain trial types within and across those blocks did fall below 75%. In particular, the trials in which "Dislike" was the evaluative stimulus exhibited significantly greater frequencies of errors than did those trial types with the

"Like" evaluative stimulus. As a result, a participant whose level of correct responding on certain trial types was well below the 75% criterion cutoff would still be included in analyses, given that the overall level of correct responding was still sufficient.

As previously discussed, the MT-IRAP handles errors in a different manner. In particular, the MT-IRAP allows more than just a single opportunity to respond to each combination of evaluative and target stimuli in each block of trials, such that if an incorrect or error response is emitted in the presence of a certain combination of stimuli, then those stimuli are presented again in a later trial within the same block, allowing the opportunity for a correct response to be emitted and associated error-free latency to be recorded. MT-IRAP data in Experiment III also indicated that, in general, significantly more error responses were emitted in the presence of the "Dislike" evaluative stimulus, however, in each case each participant was able to respond to each combination of stimuli until a fast, accurate response which complied with the implicit responding requirements was emitted. Given that error trials in the traditional IRAP involve first the emission of the incorrect key-press response, followed by the observing response associated with the "red X" which appears on screen, which is then followed by the emission of the other, correct key-press response, there is an inevitable inflation of response latencies associated with these trials, which cannot accurately be parsed out from the response latency which would have occurred if the initial key-press response had been a correct one instead. Due to this confounding influence on response latencies, the method by which the MT-IRAP handles error trials is seen as more appropriate for a latency-based assessment tool and, for this reason, is likely to be at least one factor which accounts for the ability of the MT-IRAP to better predict other behavioral measures of interest.

Another important empirical observation from the present study was that many stimuli which were included in the same conceptual category of “hard work” or “easy work”, based on topographical and semantic similarity, turned out not to be functionally equivalent, as measured by the IRAP assessments. Based on this observation, it is reasonable to conclude that IRAP scores pertaining to these lists of stimuli (i.e., “hard work” and “easy work”) would not lend themselves to predictive utility, unless a majority of the stimuli in each list were functionally related to the behaviors and associated measures to be predicted. In the case of Experiment III, the lists of “hard” and “easy” work stimuli did not correlate with or predict the behavioral persistence measure from the analog work task for either IRAP assessment. This was perhaps to be expected, given that the majority of stimuli utilized in the assessments evaluated attitudes toward liking or disliking work that is stressful, or easy, or demanding, etc., rather than persisting at work, in particular.

It was found to be the case that with both IRAP instruments the IRAP scores for the Like→Persistence trials positively correlated with and therefore to at least some extent predicted the behavioral persistence measure of number of work sessions completed. Although only the IRAP scores from the MT-IRAP were found to yield a strong predictive effect which was statistically significant, this finding from both of the IRAPs emphasizes the importance of recognizing that semantically related stimuli are not necessarily functionally related and, therefore, not all stimuli should be considered appropriate for analysis in the case of determining or trying to occasion predictive validity. Given this assumption, the method of handling error responses is paramount, as there is less room to tolerate and work with error-related latencies for individual stimuli, since there are so few responses and data available for the analysis of any particular individual stimulus.

Another difference between the two IRAP instruments which likely contributed to the differential capability of each to predict other behavior of interest is related to the additional layer of relational responding required by the MT-IRAP. Specifically, the MT-IRAP required participants to engage in an additional relational response, as cued by the contextual cue of the Truth/Lie label stimuli. To the extent that implicit and explicit relational responding is distinguished and defined by the relative extent of extended, elaborated relational responding, the MT-IRAP appears to measure relational responding which is at least one degree closer to explicit on the implicit/explicit continuum than does the traditional IRAP. Although responses on both instruments are emitted under time-pressured conditions (e.g., less than 3000 or 2000 ms), which is generally regarded as the primary procedural factor that drives measures of implicit responding, there is necessarily at least one additional relational response which must occur in the presence of the additional stimuli in the MT-IRAP, relative to the traditional IRAP. Therefore, it may be argued that the MT-IRAP measures responses that are somewhat more explicit than does the traditional IRAP.

In conjunction with this assertion, the nature of the behavioral measures which are to be predicted by an IRAP assessment may require consideration as well. For example, in the case of Experiment III, the primary behavioral measure to be predicted was that of work sessions completed, which was determined by the choice of the participant to quit the work task. This decision to quit the work task, as evidenced by the choice response of clicking the “Quit” button in the work task, did not occur under time-pressured conditions. Instead, participants were allowed as little or as much time as they needed to decide when they would ultimately quit the task. Due to this aspect of the procedure, it is possible that the MT-IRAP, which requires slightly more extended and elaborated relational responding, may yield results which are more predictive of a behavioral task, when that behavioral task occurs under conditions which are considered explicit, i.e., they involve elaborated relational responding and do not occur under heavy time pressure. Similarly, it is possible that when the behavior-to-be-predicted occurs under conditions which correspond with implicit responding, i.e., behavior occurring under high time pressure in which extended, elaborated relational responding is not possible, the traditional IRAP may yield data which are more predictive of the behavior of interest. However, even if the foregoing hypothesis were an accurate description of the phenomena in question and the IRAP were theoretically to have a predictive advantage in some cases—depending upon the behavior-to-be-predicted—it would still suffer from the mishandling of error responses, as detailed above. The extent to which this procedural difference between the two IRAPs does in fact influence the predictive capability of either is of course only speculation at this time; however, this line of questioning is amenable to empirical, experimental investigation and can therefore be effectively answered through future research.

That there was a clear sequence effect relating to predictive validity for the persistence analyses raises the obvious question of the nature of this influence on participant responding during the IRAP assessments. Given that for this measure in particular there was better performance observed on the first IRAP in the sequence, in terms of correlation with the persistence work task measure, it is possible that fatigue was a factor. This notion is possibly supported by the observation that the MT-IRAP is widely acknowledged, albeit anecdotally, as the much more difficult and taxing of the two instruments; therefore, the fact that performance was hampered to a greater extent when the traditional IRAP followed the MT-IRAP rather than vice versa, suggests that fatigue could have played a role.

The foregoing interpretation is challenged somewhat by the same sequence effect analysis pertaining to the list-level IRAP scores for Like→Hard work stimuli. In this particular analysis, virtually the opposite effect was observed for each of the instruments, which suggests that fatigue was possible not an influence. By contrast, fatigue could still have been a factor, however, the effects of fatigue on each individual stimulus could have been largely washed out or masked, due to the analysis being conducted at the list level. Any influences on participant responding by factors such as fatigue may be more easily detected at the level of the individual stimulus; this, however, requires additional research to determine. This finding also raises the broader question of the quality of IRAP data under differing circumstances, such as when a participant is already tired or fatigued before beginning the assessment, or how long an IRAP can be before fatigue sets in, or how many IRAPs can be completed by a participant within a given period of time. Also of particular interest is whether the fact that the two IRAPs in the present study used the same target stimuli is at issue, and whether completing two different IRAPs in succession is more or less susceptible to the influences which were at play in the present study. Again, future research will have to parse out and directly test each of these questions.

The general lack of convergence between the explicit attitude measures and the implicit IRAP measures is interesting. Implicit measures tend to diverge from explicit measures when the attitudes being assessed are of a socially sensitive nature (Barnes-Holmes, Barnes-Holmes, et al., 2010), however, there is no objective means of determining which topics are socially sensitive, thus it is not possible to say whether the concepts and attitudes assessed in the present study were.

There is some evidence to suggest that perhaps some of the specific stimuli utilized in the present study may have been socially sensitive to some extent, or at least susceptible to broad socio-cultural norms, which may have evoked strong conventional responses on the part of participants. For example, the explicit survey item with the second greatest average score across participants (2.42) was for the item pertaining to the stimulus “Lazy”. No participants answered that they are sometimes lazy at work; however, the IRAP results did not indicate the similar biases at the implicit level, as more than half of all participants indicated a Like→Lazy bias on at least one of the IRAP assessments. A similar pattern was observed for the explicit survey item asking participants about “giving up” (question #4), to which all but one participant indicated to at least some extent that they do not give up. The average survey score for this item was 2.05, which was among the strongest positive scores for all items. Again, half of all participants indicated a pro-Quitting implicit bias on at least one of the two IRAP instruments. For items such as these, it is likely that participants explicitly answered that they are not lazy and do not give up (i.e., quit), based on the general social stigma and disapproval related to such descriptions of people and their actions. It is likely that even people who are “lazy” and “quitters”, based on objective, behavioral operational definitions of such adjectives, would not describe themselves as such and openly admit such attitudes to others. In this regard, these concepts may be said to be socially sensitive.

While the explicit scores for the Persistence stimuli did not provide any predictive utility in terms of predicting number of sessions completed at the analog work task, the overall survey scores for all stimuli did demonstrate a modest amount of predictive validity with respect to the performance metric of average correct trials completed. More research is clearly needed to determine the conditions under which explicit attitude measures predict other behaviors of interest, under which conditions explicit measures are susceptible to social sensitivity and self-presentational influences, and under which conditions either explicit or implicit measures are the better predictor of behaviors of interest.

General Discussion

The three experiments in the present study attempted to begin to systematically answer some of the most important questions currently surrounding emerging research in the area of implicit attitudes and biases and their place in a behavioral account of human behavior. In recent years two different IRAP instruments have been developed and to the extent they theoretically measure the same aspects of complex human behavior, such as language and cognition, it is important to empirically investigate the extent to which they appear to similarly measure that which they claim to and, if at all possible, which instrument appears to be more effective in its aims. While not definitive, Experiment I of the present study suggests that the two instruments do not reliably (i.e., approximately half of the time) yield similar results on a within-subject basis. This in itself is a concern, as it raises some questions about the extent to which the phenomena of interest, i.e., implicit attitudes and cognitions, can be reliably measured by such instruments. Assuming that they can be reliably measured, the next logical question involves asking which of the two instruments produces more reliable and accurate results. The results of Experiment I suggest that the MT-IRAP produces more consistent, reliable results on a within-subject basis, however, Experiment I was not able to shed light on which of the two instruments provides more accurate results, especially since “accurate with respect to what?” is difficult to answer.

Experiment II attempted to come one step closer to determining which of the two IRAP instruments provides more accurate results by identifying some behavioral measures of more overt, easily observable patterns of behavior, pertaining particularly to an organizational workplace setting, which could be measured and correlated with the IRAP results. While there are likely an infinite number of measures across a myriad of settings which could conceivably be utilized for such an analysis, the present study represented an important initial attempt (in addition to Nicholson & Barnes-Holmes, 2012) to do so. The results of Experiment II suggest that both IRAPs can to at least some extent predict how an individual is likely to behave, at least as far as the particular stimuli and behavioral measures in Experiment II are concerned, and that the MT-IRAP appears to have an advantage over the traditional IRAP in doing so.

Experiment III emerged as the primary investigation among the present set of studies, as it not only provided a replication of Experiment I with a greater number of participants, but it also provided more power in terms of answering the question of predictive validity addressed in Experiment II, as each participant was exposed to both

IRAP instruments in Experiment III, thus allowing the comparisons to be made with the same pool of participants. The results of Experiment III clearly indicated that the specific stimulus combination of Like→Persistence correlated to at least some extent with persistence at the analog organizational work task, however, only the MT-IRAP scores correlated significantly with persistence at the work task and demonstrated a strong effect, in terms of predictive validity. Analyses of IRAP scores at the level of lists of stimuli did not yield any significant correlations with any of the work task measures, although the MT-IRAP scores for the Trial Type of Like→Hard Work words (Trial Type 1) correlated somewhat and approached significance with average trials correct per session during the work task, which was construed as the most general measure of performance during the work task.

The results of Experiment III are somewhat consistent with those of Nicholson and Barnes-Holmes (2012) and Carpenter et al. (2012), in that IRAP results were to at least some extent generally predictive of another behavior of interest in a particular context. However, the results of Experiment III differed from these other studies in that only the IRAP scores for certain stimuli, in this case only the "Persistence" stimulus, were predictive of the behavior in question. This is almost certainly at least partly due to the fact that Nicholson and Barnes-Holmes utilized pictures for target stimuli in their study, all of which were formally very similar (i.e., different pictures of spiders), while the present study used written words for target stimuli, which were only loosely semantically related and shown not to be functionally equivalent for many participants.

Interestingly, Carpenter et al. utilized target stimuli in the form of written words and phrases, as did the present study, and it is possible that those target phrases were more functionally equivalent than those in the present study, given they were more directly tied to the conceptual attitudes in question (i.e., consequences of cocaine use), than were the attitudes in the present study (various attitudes toward broad work-related concepts). In addition, only the MT-IRAP significantly predicted persistence at the work task in Experiment III, whereas the traditional IRAP did not, and this again is not entirely consistent with the findings of Nicholson and Barnes-Holmes or Carpenter et al. As discussed previously, this may be due to a number of factors, including but not limited to the different procedures for handling incorrect error responses within each assessment, as well as the fact that the MT-IRAP requires slightly more explicit responding than does the traditional IRAP, which may or may not affect its ability to predict other behaviors, depending upon whether those behaviors occur under conditions that promote more implicit or explicit responding along that continuum. It is possible that the avoidance of spiders, as measured in Nicholson and Barnes-Holmes, was a response that would be considered more implicit, in terms of emitted more quickly and consisting of fewer elaborated relational responses, than was the quitting of the work task in Experiment III, which could in part account for the success of the traditional IRAP to predict spider avoidance in that study. The same could possibly be argued for the primary behavioral measure in the Carpenter et al. study, namely, cocaine use, as the particular choice of whether to engage in drug use in that moment may occur under greater time and other pressures than did the choice to quit the work task in the present study.

Nicholson and Barnes-Holmes (2012) and Carpenter et al. (2012) represented the first studies of their kind, in terms of evaluating the ability of the IRAP to predict other behavior patterns. As one of the only other studies to investigate that question, the present study replicates and extends those findings to a certain extent, based on various factors. A notable difference between the studies is the fact that the implicit attitudes and overt behaviors of Nicholson and Barnes-Holmes were very specific, indeed being conceptually tied to instances of human psychopathology, while the present study incorporated implicit attitudes and behavioral measures that are somewhat more generally applicable to the broader population. Similarly, the Carpenter et al. study utilized a very specific population, cocaine-dependent participants, and very specific behavioral measures, such as treatment attendance and outcomes. While the present study still employed an analog work task in a laboratory setting, which is not as generalizable as a naturalistic organizational setting would have been, it offered an incremental improvement in terms of demonstrating the applicability of IRAP assessments, particularly the MT-IRAP in the present case, to other areas of human functioning. Future studies will have to devise other means of measuring relevant patterns of overt behavior in various settings, in order to better evaluate the effectiveness of any IRAP instrument when predicting behaviors of interest in those settings.

The interpretation of why the MT-IRAP was better able than the traditional IRAP to predict the behavior of interest in this study, and why only certain IRAP scores appeared to be relevant to such predictive capability, is tied to a number of various conclusions which were drawn from the empirical observations made during the current series of experiments. These observations and conclusions they support are not only relevant to the present studies, but are broadly applicable to IRAP research in general. The conceptual conclusions drawn from this series of experiments have already been discussed in detail (see Experiment III Discussion) and will only be briefly mentioned here. A primary concern among the differences in the two instruments is the procedure for handling error responses, which in the case of the traditional IRAP involves keeping the necessarily inflated error latencies for the purpose of analysis. Additionally, it was observed that many of the stimuli which were assumed to participate in relations of coordination with another, for example, the "hard work" words, were found to have varying stimulus functions for each participant, as measured by the IRAP assessments. This was the case for both IRAPs, however, the traditional IRAP may suffer more from the unknown variability in stimulus function among target stimuli, given that its results are always provided at the list level and, further, results are usually analyzed by groups of participants, which can be problematic if the participants respond differently to each of the different stimuli assumed to be functionally equivalent within a given list of stimuli. (See Levin et al. for a review of these conceptual concerns.) Lastly, it has also been noted that the MT-IRAP requires an additional relational response on the part of the participant, due to the Truth/Lie contextual cue presented with the evaluative and target stimuli. While this aspect of the MT-IRAP can be argued to involve slightly more explicit responding on the part of participants, it remains to be determined whether this particular aspect of the assessment influences its ability to predict other behavioral measures under certain conditions.

Limitations. There are a number of limitations associated with the present set of studies, as follows. Firstly, it became clear early on that the choice of evaluative stimuli requires greater consideration than may have been previously thought. For instance, in the case of the present research, it was assumed that "Like" and "Dislike" were two clear, straightforward evaluative terms with which participants would be familiar and fluent and able to respond effectively. It also appeared that these stimuli would work well together for the purpose of this research, in that Dislike was merely the opposite of Like, as denoted by the negative prefix dis-. However, the results showed that the vast majority of participants emitted significantly more errors on Dislike trials than on Like trials. The explanation of this finding is based on two conclusions. The first is that people are typically more verbally fluent in relationally responding in terms of what they like (e.g., I like this; I don't like that), whereas they are much less fluent in responding in terms of what they dislike (e.g., I dislike this, I don't dislike that). The latter example exemplifies the second conclusion drawn from these observations: utilizing evaluative stimuli which include a negative prefix (e.g., dis-, un-) can be problematic and should be avoided, especially when using the MT-IRAP, which also includes the Lie trial label on some trials (in such cases, depending upon the target stimulus, there can be up to a triple negative in a single trial, which is incredibly difficult and frustrating to deal with and will undoubtedly infuriate the participant, which could lead to other research problems).

The second concerns the choice of target stimuli to include in the IRAP assessments. As mentioned earlier, the typical procedure is to merely choose words that seem to be semantically related and have some amount of face validity, in terms of their functional equivalence. The results of the present experiments indicated that this was not a sufficient means of choosing target stimuli, especially if the IRAP analyses are to be conducted at the level of the list of stimuli, as is typically done with the traditional IRAP. One can get away with this a little more with the MT-IRAP, in which the emphasis is on individual stimuli with the assumption that the target stimuli are not necessarily functionally equivalent, however, depending upon the goals of the analyst, it may be desirable to have as many functionally equivalent stimuli used in the assessment as possible, rather than find out after the fact that there is only one or few viable (in terms of stimulus function) stimuli to use for analytical purposes.

Another limitation of the present study centers around the objectively observable measure-to-be-predicted by the IRAP assessments. This is no doubt the most challenging aspect of research like that of the present experiments, as it is quite difficult to identify a behavioral measure which corresponds closely (and presumably functionally) with the stimuli being assessed by an IRAP. Such behavioral measures must be amenable to overt, objective observation and measurement and also capable of being measured in a setting of relative convenience for the researcher. As an example, if a researcher wanted to identify whether someone who completed a race/ethnicity IRAP was in fact racist or discriminatory toward a certain racial/ethnic group, as may be indicated by an IRAP, how could the researcher make repeated, objective observations to utilize for that analysis? It is difficult to imagine how that study would look, though it is surely not impossible to conduct. In the case of the present study, an analog, pay-for-performance work task in which participants could choose to quite at their discretion was employed as a means to measure persistence at work. While this was a first attempt at such research and the present study did yield positive results, suggesting that this task was at least sufficient for the present research question, there is no doubt that a better, more naturalistic behavioral measure could have been developed in order to test the predictive validity of the IRAP instruments. Of course, the behavioral measure-to-be-predicted must vary with the concept(s) being assessed by the IRAPs, which will continue to make such studies difficult to develop and implement.

There are other limitations associated with the present set of studies, for example, that the study was conducted in a laboratory and the work task was an analog to the naturalistic workplace setting. This is particularly important when considering the generalizability of the present findings to the organizational setting. Associated with this are a number of other factors which limit the findings of the present study, including the motivation of participants, in terms of behavioral contingencies, to perform and persist at the task, the pay-for-performance, piece-rate nature of the monetary compensation, and the fidelity of the analog work task and the extent to which it approximates job requirements in naturalistic settings. With respect to the latter, the increasing demands in the present study were implemented in order to prevent participants from persisting indefinitely at the task, as well as to occasion potentially stressful conditions which are presumed to exist in naturalistic work settings, however, the ability of the analog work task to approximate that aspect of organizational settings was limited as well. Lastly, a limitation which is often ubiquitous in much research pertains to the participant pool, namely, undergraduate freshman psychology students.

Future research. Future research in this area is wide open and a number of future studies have already been suggested, based on the observations made in the present study and the observed limitations of the present study. For example, future IRAP research should focus on methods by which to identify evaluative stimuli that will assess the attitudes of interest without being too difficult to respond to, thereby occasioning fewer error responses.

Future studies should also address potential *a priori* means of selecting target stimuli for an IRAP assessment, to the extent that a certain composition of target stimuli are needed (e.g., many functionally equivalent or similar stimuli). IRAP research should also begin considering how to work with analyses conducted at the level of individual stimuli, given that there are so few response latency data available for any given individual stimulus. A balance will have to be struck between collecting enough latency data to provide for a confident analysis of any individual stimulus and not extending the number of trials and duration of an IRAP assessment an untenable amount.

Additional research is needed to replicate the findings of the present series of studies, as well as to extend these initial findings. In particular, more creative ways of developing meaningful behavioral measures to correlate with IRAP results is necessary. The findings of the present studies are only a first step and are still limited to a very specific set of IRAP stimuli and particular analog work task measure. As such, the external validity and generalizability of these findings is limited for the time being, pending further research aimed at broadening them. Given the relatively recent development of such implicit measures as the IRAP, there are a limitless variety of studies which can be designed to do just that.

As additional research with the various iterations of the IRAP is conducted, there will be important implications to be considered, especially concerning the use of IRAP assessments in applied settings. To date, there has been no published research employing an IRAP in an applied setting, as thus far the research has focused on more basic research questions, such as whether the IRAP possesses any predictive utility under controlled laboratory conditions, as in the present study and other recent work (e.g., Nicholson and Barnes-Holmes). However, as more research corroborates the findings of Nicholson and Barnes-Holmes and the present study, the potential application of the IRAP will become increasingly appealing in various contexts, for instance, for selection and placement purposes within organizations. Applied research in these settings will be of utmost importance, in order to ensure that IRAPs and other similar assessments are utilized appropriately, based on empirical evidence. The same challenges to this research will apply, in terms of aligning target concepts assessed by an IRAP with the behaviors and outcomes of interest in the applied setting.

Some primary examples of future applications of IRAP assessments include educational, as well as organizational settings, in which an ever increasing importance is placed on cultural diversity and cultural competence, not only at the levels of students and employees, but faculty, administrators, managers, and executives as well. Such attitudes toward cultural and ethnic concepts have been some of the most highly targeted within the domain of implicit attitudes research and will be of continued interest in the context of ongoing economic and industrial globalization. Interest in other social biases, such as attitudes toward overweight and obese individuals, will provide a strong niche for IRAP assessments among health care providers, for example, given the evolving landscape of health care and the health of individuals in this country. Interestingly, there is already a growing literature pertaining to this very topic in medical education journals, utilizing the more popular IAT assessment (see Puhl & Heuer, 2009, for a recent review). It is important to note that the popularity and demand of the IAT assessment in medical science literature has increased without any systematic analysis of its predictive utility. Given the theoretical and procedural superiority of the IRAP relative to the IAT, there is great potential for future research and application within this particular area, as well as beyond.

In conclusion, it should be reemphasized that this study represents a very initial step in the continual process of attempting to establish the extent to which an IRAP assessment can predict how an individual may behave in other naturalistic settings of interest. The present study consists of a very small, specific, and contrived sample of the many different implicit attitudes which may be of interest (e.g., attitudes about work) and the very many complex contexts in which people behave on a daily basis. Whether the stimuli utilized for the IRAP assessments in the present study are the best sample for assessing attitudes towards the workplace cannot yet be answered. Similarly, whether the analog data entry work task incorporated in the present study is an ideal analog contrivance to represent the naturalistic workplace setting is also unknown at this time. It goes without saying that studies attempting to answer the primary question of the present research must continue to be conducted using a variety of different IRAP assessments and stimuli as well as overt patterns of behavior with which to correlate those IRAP results. Researchers will have to be very creative with respect to finding ways to measure human behavior in the settings of interest, in order to occasion a better understanding of the conditions under which an IRAP assessment can be utilized to predict with some amount of certainty other behavioral patterns of interest.

References

Barnes-Holmes, D., Barnes-Holmes, Y., Power, P., Hayden, E., Milne, R., & Stewart, I. (2006). Do you really know what you believe? Developing the Implicit Relational Assessment Procedure (IRAP) as a direct measure of implicit beliefs. *The Irish Psychologist, 32,* 169-177.

Barnes-Holmes, D., Barnes-Holmes, Y., Stewart, I., & Boles, S. (2010). A sketch of the Implicit Relational Assessment Procedure (IRAP) and the relational elaboration and coherence (REC) model. *The Psychological Record, 60,* 527-542.

Barnes-Holmes, D., Murtagh, L., Barnes-Holmes, Y., & Stewart, I. (2010). Using the Implicit Association Test and the Implicit Relational Assessment Procedure to measure attitudes towards meat and vegetables in vegetarians and meat-eaters. *The Psychological Record, 60,* 287-305.

Barnes-Holmes, D., Waldron, D., Barnes-Holmes, Y., & Stewart, I. (2009). Testing the validity of the Implicit Relational Assessment Procedure and the Implicit Association Test: Measuring attitudes toward Dublin and country life in Ireland. *The Psychological Record, 58,* 389-406.

Bryant, E. C. (1960). *Statistical Analysis.* York, PA: Maple Press Company.

Carpenter, K. M., Martinez, D., Vadhan, N. L., Barnes-Holmes, D., & Nunes, E. V. (2012). Measures of attentional bias and relational responding are associated with behavioral treatment outcome for cocaine dependence. *The American Journal of Drug and Alcohol Abuse, 38(2),* 146-154.

De Houwer, J. (2002). The Implicit Association Test as a tool for measuring dysfunctional associations in psychopathology: Strengths and limitations. *Journal of Behavior Therapy and Experimental Psychiatry, 33,* 115-133.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem and stereotypes. *Psychological Review, 102,* 4-27.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 78,* 1464-1480.

Harmon, K., Strong, R., & Pasnak, R. (1982). Relational responses in tests of transposition with rhesus monkeys. *Learning and Motivation, 13,* 495-504.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame Theory: A post-Skinnerian account of human language and cognition.* New York: Kluwer/Plenum.

Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record, 61,* 465-496.

Kantor, J. R. (1958). *Interbehavioral psychology: A sample of scientific system construction.* Bloomington, IN: Principia Press.

Levin, M. E., Hayes, S. C., & Waltz, T. (2010). Creating an implicit measure of cognition more suited to applied research: A test of the Mixed Trial-Implicit Relational Assessment Procedure (MT-IRAP). *International Journal of Behavioral Consultation and Therapy, 6(3),* 245-262.

Nicholson, E., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure (IRAP) as a measure of spider fear. *The Psychological Record, 62,* 263-278.

Puhl, R. M., & Heuer, C. A. (2009). The stigma of obesity: A review and update. *Obesity, 17(5),* 941-964.

Salkind, N. J., & Green, S. (2011). *SPSS Quick Starts.* Upper Saddle River, NJ: Prentice Hall.

Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *The Quarterly Journal of Experimental Psychology, 60(3),* 291-309.

Sidman, M. (1986). Functional analysis of emergent verbal classes. In T. Thompson & M. E. Zeiler (Eds.), Analysis and Integration of Behavioral Units (pp. 213-245). Hillsdale, N.J.: Laurence Erlbaum Associates.

Sidman M., & Tailby, W. (1982). Conditional discrimination versus matching to sample: An expansion of the testing paradigm. *Journal of the Experimental Analysis of Behavior, 37,* 5-22.

Tables

Table 1.
Target, Evaluative, and Response Stimuli for Both IRAPs

Hard work stimuli	Easy work stimuli	Evaluative stimuli	Response options	Practice stimuli[a]
Effortful	Lazy	Like	Yes	Flower
Difficult	Easy	Dislike	No	Rose
Demanding	Slow			Daisy
Hard	Leisurely			Tulip
Stressful	Relaxing			Spider
Complex	Simple			Cockroach
Busy	Calm			Beetle
Persistence	Quitting			Hornet

[a] Practice stimuli for MT-IRAP only. Practice for traditional IRAP included the same hard and easy work stimuli used for testing.

Table 2.
Pearson's Product-Moment Correlation Coefficient for All Trial-Types Across Both IRAPs

Participant	*r*
PP1-1	.85
PP1-2	.29
PP1-4	-.11
PP1-5	.18
PP1-6	.59
PP1-7	.64
PP1-8	.92
PP1-9	.63
PP1-11	-.16
PP1-12	.18
All	.19

Table 3.
Pearson's Product-Moment Correlation Coefficient for All Trial-Types Across Both IRAPs with Errors Removed from Traditional IRAP

Participant	*r*
PP1-1	.71
PP1-2	.30
PP1-4	-.35
PP1-5	.30
PP1-6	.69
PP1-7	.87
PP1-8	.92
PP1-9	.59
PP1-11	-.29
PP1-12	.46
All	0.23

Table 4.
Pearson's Product-Moment Correlation Coefficient r for Individual Stimuli Across Both IRAPs

Participant	All trial types	Like trial types	Dislike trial types
PP1-1	.10	.06	-.03
PP1-2	-.21	-.24	-.21
PP1-4	.02	.07	.32
PP1-5	-.02	-.25	.09
PP1-6	.05	.01	.14
PP1-7	.41	.46	.40
PP1-8	.25	.26	.39
PP1-9	.67	.75	.48
PP1-11	-.11	-.01	-.14
PP1-12	.00	-.03	.11
All	.14	.16	.12

Table 5.
Work Task Performance Measures in Experiment II

Participant	No. sessions	Avg. correct/ session	Total % correct	Failed to meet min.	No. sessions after fail	Revenue earned
PP2-1	4	9.75	95%	No	-	$1.17
PP2-2	25	16.3	96%	No	-	$11.76
PP2-3	23	14.6	96%	Yes	3	$8.22
PP2-4	19	4.1	25%	Yes	18	$0.00
PP2-5	11	10.7	86%	Yes	4	$2.19

Table 6.
Pearson's Product-Moment Correlation Coefficient for All Trial-Types Across Both IRAPs in Experiment III

Participant	*r*	Participant	*r*
P1	-.91	P19	-.67
P2	-.60	P20	.85
P3	-.66	P21	.63
P4	.20	P22	.15
P5	.59	P24	.51
P8	.27	P25	.59
P10	.38	P28	-.03
P11	.26	P29	.81
P12	-.63	P30	-.56
P13	.21	P32	-.46
P17	-.92	P33	-.48
All	**-.01**		

Table 7.
Pearson's Product-Moment Correlation Coefficient r for Individual Stimuli Across Both IRAPs in Experiment III

Participant	All Trial Types	Like Trial Types	Dislike Trial Types	Participant	All Trial Types	Like Trial Types	Dislike Trial Types
P1	.04	.17	-.04	P19	-.07	-.02	-.41
P2	-.03	-.05	.19	P20	0.49	.57	.40
P3	-.38	-.31	-.45	P21	.06	.12	.01
P4	-.01	.45	-.23	P22	.18	.15	.21
P5	.27	.36	.10	P24	.18	.18	.14
P8	.31	.38	.17	P25	.03	.09	-.22
P10	.19	.18	.19	P28	.25	.12	.56
P11	-.27	-.34	-.23	P29	.13	-.18	.35
P12	.02	-.15	.23	P30	-.24	-.06	-.05
P13	.04	.15	-.05	P32	.09	.22	.01
P17	-.08	.27	-.33	P33	.27	.29	.24
All	**0.05**	.10	.02				

Table 8.

Work Task Performance Measures in Experiment III

Participant	No. sessions	Avg. correct/ session	Total % correct	Revenue earned
P1	24	6.4	89%	$0.00
P2	28	13.7	91%	$4.80
P3	18	11.5	54%	$1.92
P4	12	18.3	98%	$6.57
P5	3	14.7	98%	$1.32
P8	19	12.9	87%	$3.60
P10	8	6.5	93%	$0.24
P11	19	13.6	89%	$5.28
P12	19	14.1	96%	$6.03
P13	3	5.0	100%	$0.00
P17	6	8.7	87%	$0.60
P19	6	8.0	80%	$0.54
P20	13	13.6	94%	$5.31
P21	4	10.0	91%	$0.99
P22	9	12.1	97%	$3.12
P24	3	7.0	84%	$0.57
P25	23	16.4	90%	$9.36
P27	10	9.2	87%	$0.81
P28	10	11.1	92%	$2.79
P29	13	11.0	88%	$2.37
P30	36	20.2	97%	$13.50
P32	3	11.7	88%	$1.05
P33	21	16.1	96%	$9.69

Table 9.

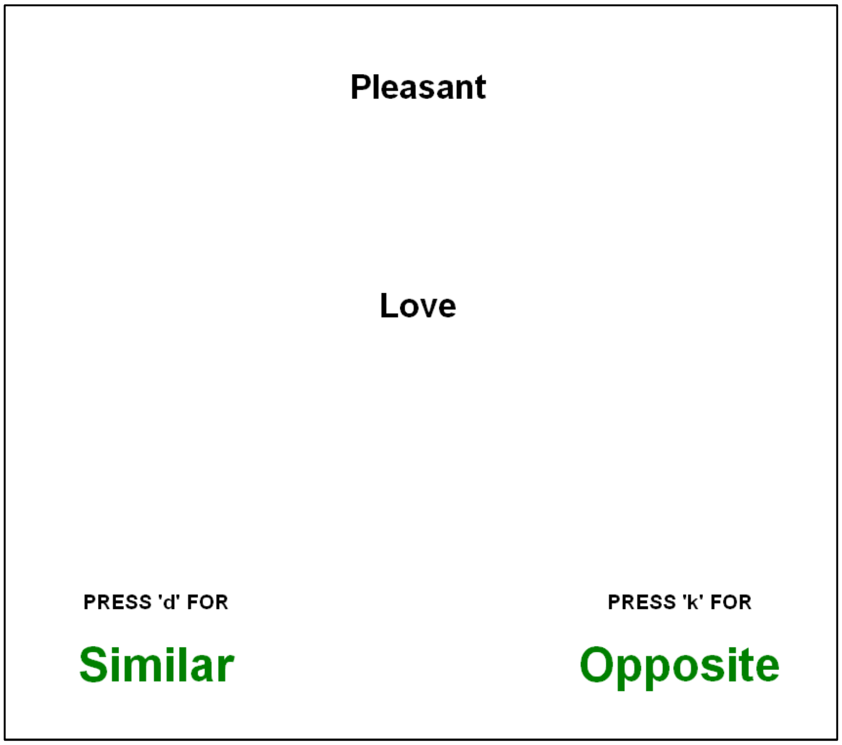
Explicit Survey Scores

Question	Avg. Score	No. Positive Scores	No. Negative Scores
1 (Easy)	0.32	10	9
2 (Effort)	2.16	19	0
3 (Difficult)	1.47	16	3
4 (Give up/Quit)	2.05	18	1
5 (Relax)	0.79	13	6
6 (Demanding)	1.68	18	1
7 (Lazy)	2.42	19	0
8 (Persist)	1.84	17	2
9 (Stressful)	0.58	13	6
10 (Easy 2)	-0.05	8	11
11 (Calm)	-1.21	2	17
12	1.21	16	3
13 (Effort 2)	2.63	19	0
14 (Complex/Simple)	2.32	18	1
All	**1.37**		

Table 10.
Pearson's Product-Moment Correlation Coefficient r for Explicit and Implicit Measures in Experiment III

Question	TR-IRAP	MT-IRAP
1 (Easy)	.02	.16
2 (Effort)	.43	.28
3 (Difficult)	-.30	-.46*
4 (Give up/Quit)	.04	.13
5 (Relax)	-.25	.28
6 (Demanding)	.00	-.50*
7 (Lazy)	.08	.41
8 (Persist)	-.07	.04
9 (Stressful)	-.32	-.34
10 (Easy 2)	.11	.46*
11 (Calm)	.09	.25
13 (Effort 2)	.00	-.19
14 (Complex)	.09	-.13
14 (Simple)	-.06	-.02
All Like-Hard work	**.12**	**.01**
All Like-Easy work	**-.20**	**-.17**

Figures



Pleasant

Love

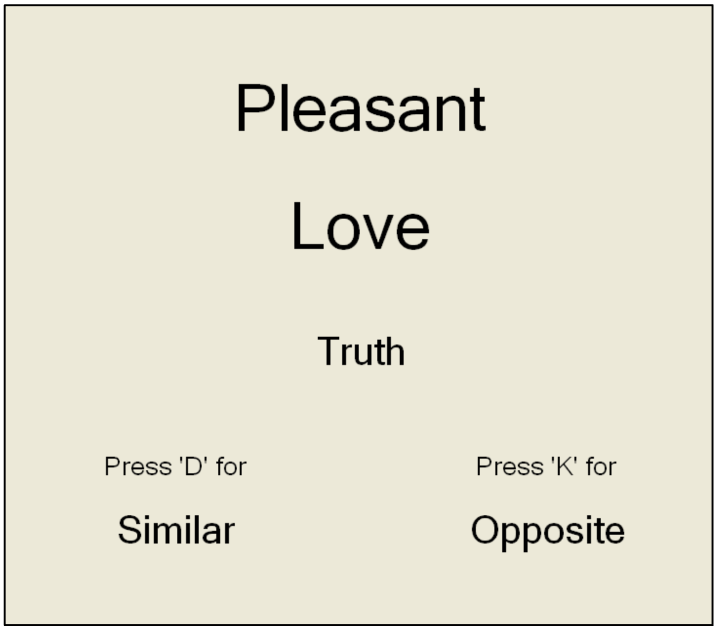
PRESS 'd' FOR

Similar

PRESS 'k' FOR

Opposite

Figure 1. Screenshot of traditional IRAP. Certain blocks (i.e., Pattern 1) require the participant to respond “Similar,” while other blocks (i.e., Pattern 2) require the participant to respond “Opposite,” and these response latencies are analyzed to determine which occurs more quickly, on average.



Pleasant

Love

Truth

Press 'D' for

Similar

Press 'K' for

Opposite

Figure 2. Screenshot of MT-IRAP. The participant can respond either “Similar” or “Opposite” in the presence of the Truth label, based on his/her explicit choice, and must therefore provide the other response (either “Similar” or “Opposite”) when these stimuli are in the presence of the Lie label.

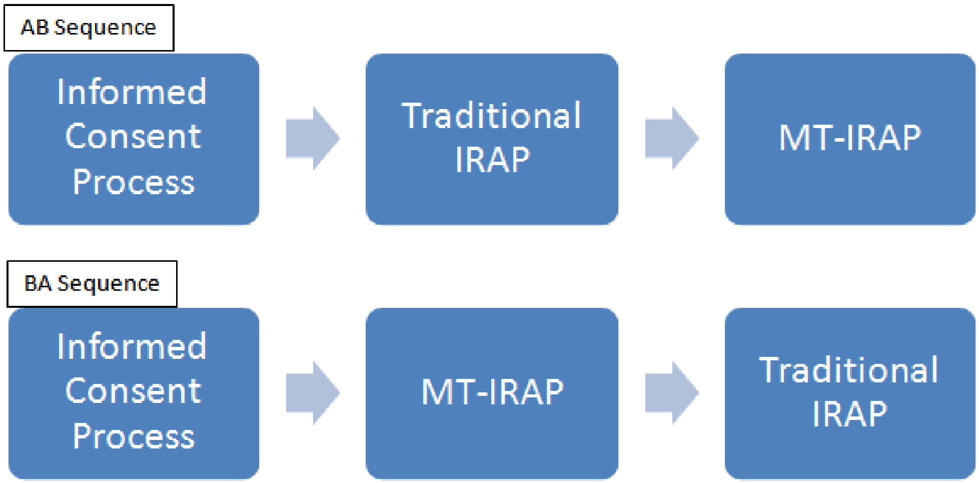


Figure 3. Visual depiction of experimental procedure for both AB and BA sequences in Experiment I.

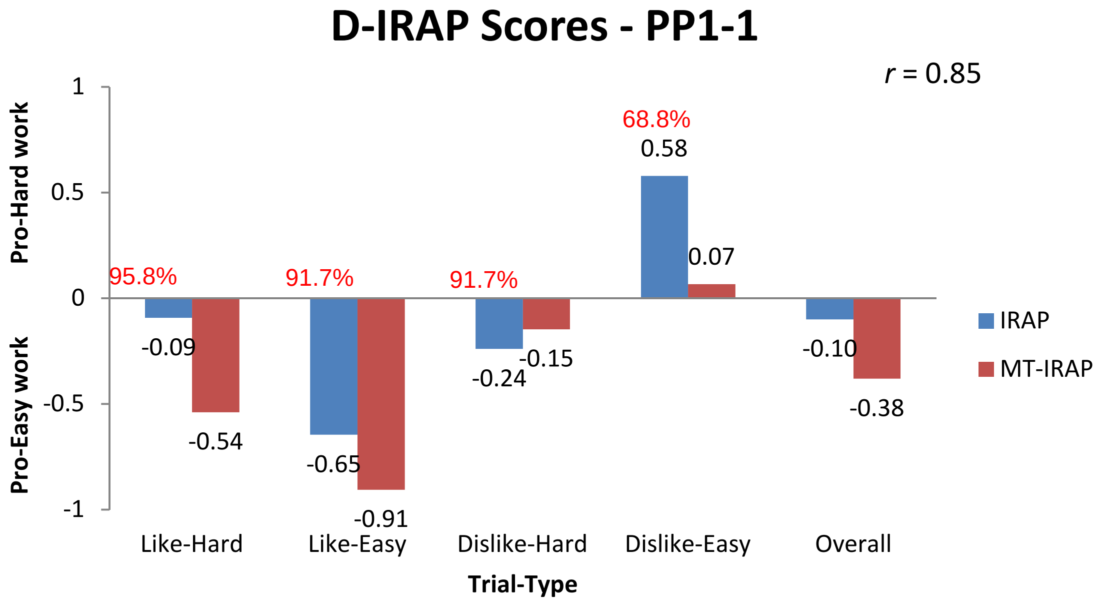


Figure 4. List-level, Trial Type D-IRAP scores for both IRAP instruments for participant PP1-1, representative of a strong positive correlation, from Experiment 1. Percentage correct shown in red for traditional IRAP only.

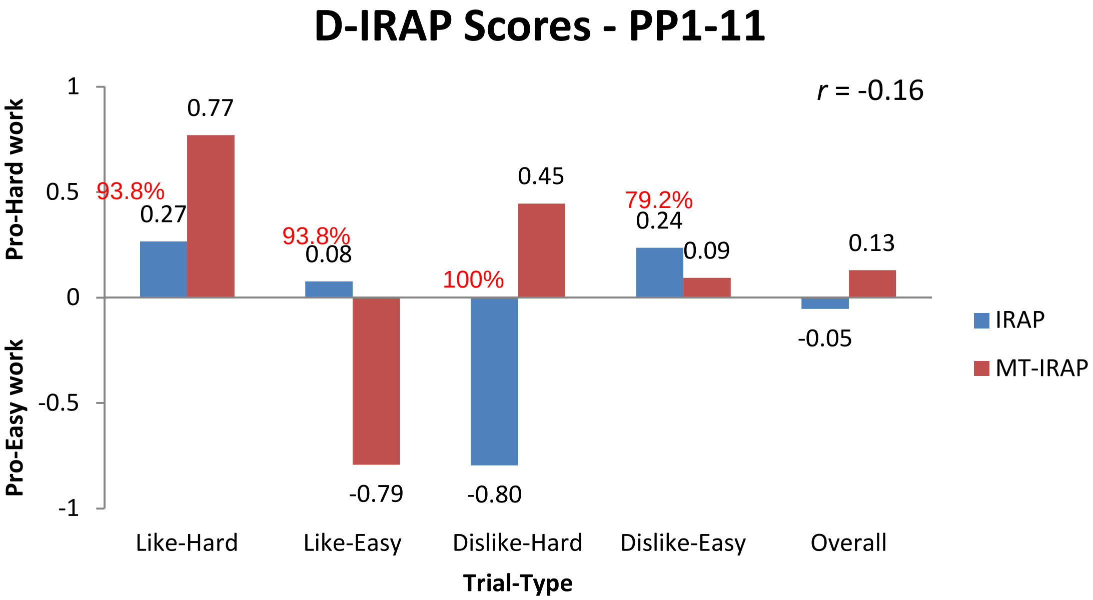


Figure 5. List-level, Trial Type D-IRAP scores for both IRAP instruments for participant PP1-11, representative of a weak negative correlation, from Experiment 1. Percentage correct shown in red for traditional IRAP only.

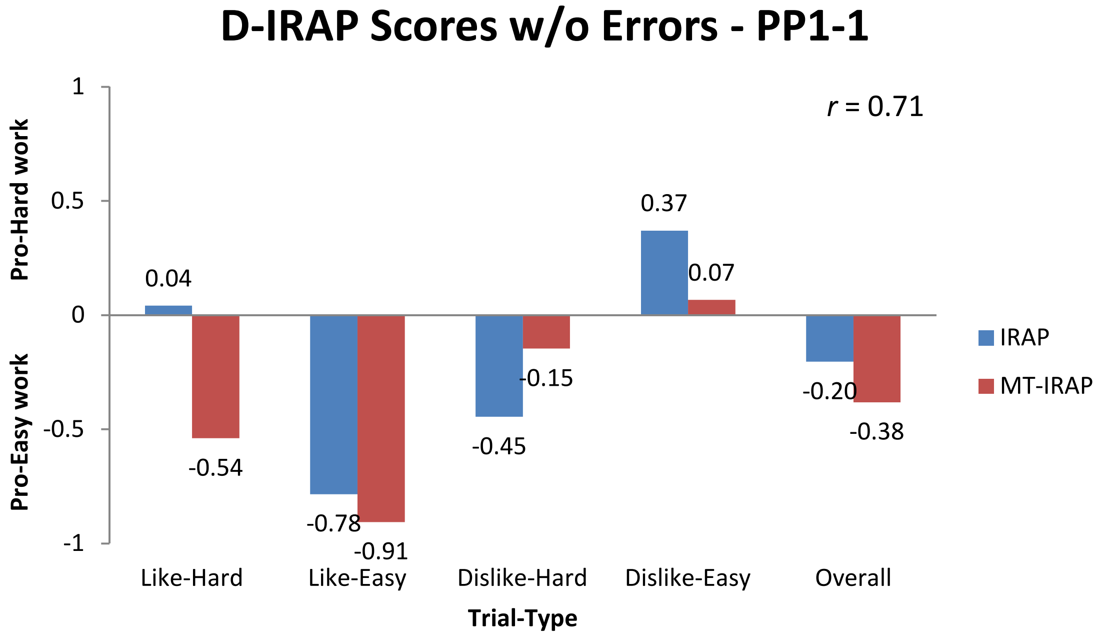


Figure 6. List-level, Trial Type D-IRAP scores for both IRAP instruments for participant PP1-1, with error responses for Traditional IRAP removed from analysis.

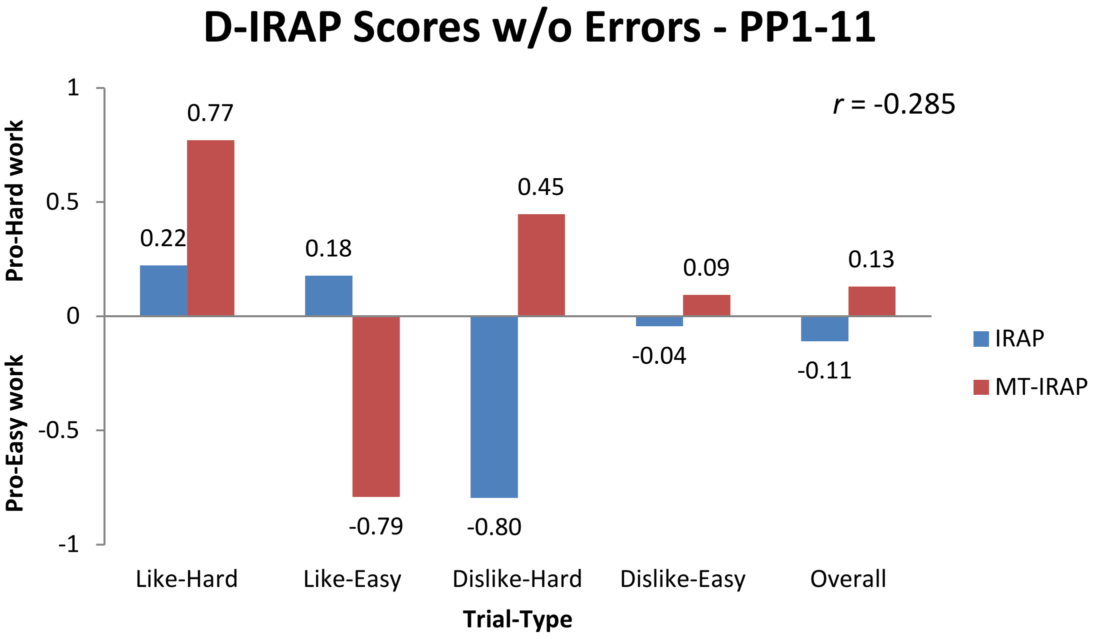


Figure 7. List-level, Trial Type D-IRAP scores for both IRAP instruments for participant PP1-11, with error responses for Traditional IRAP removed from analysis.

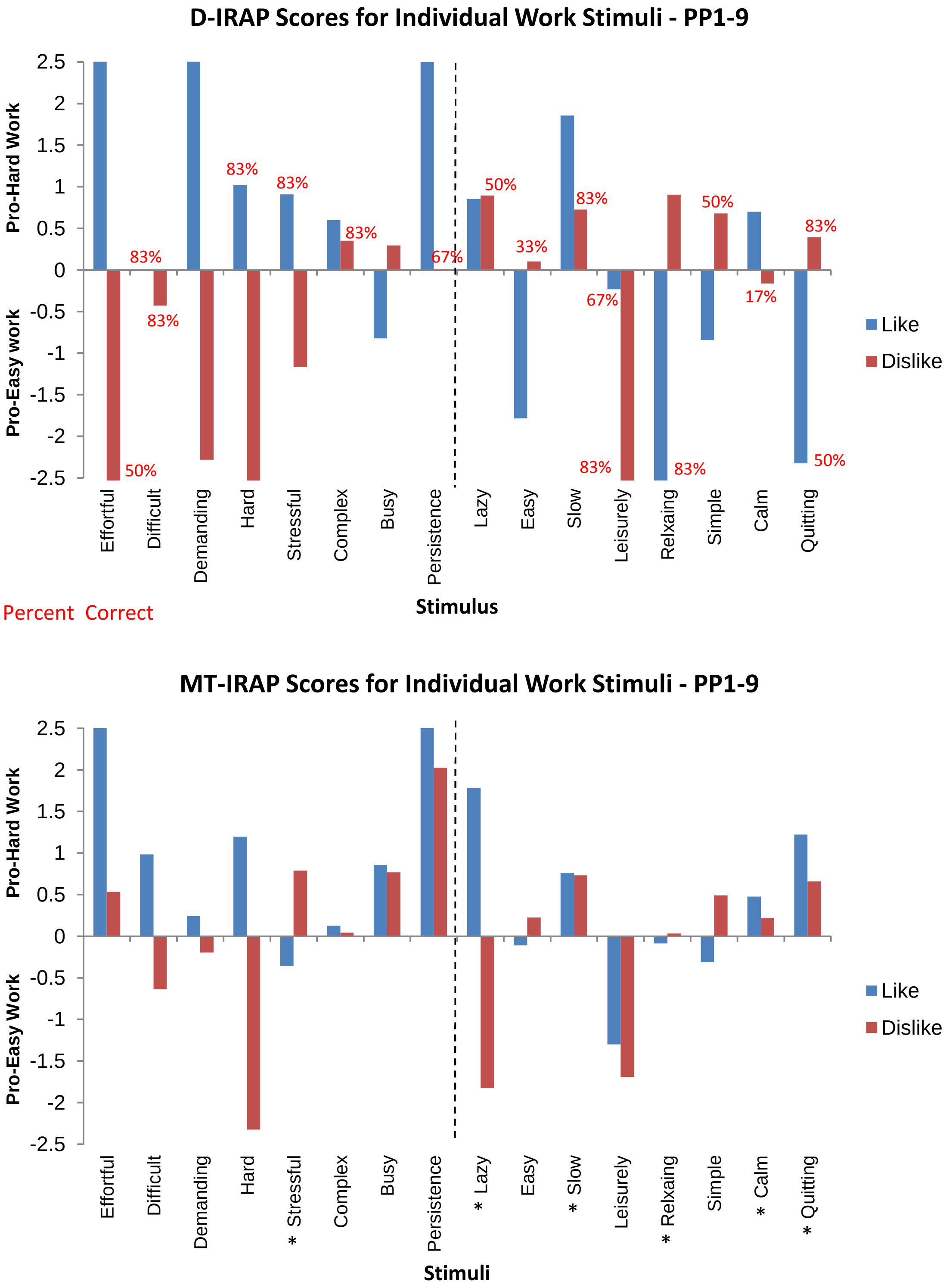
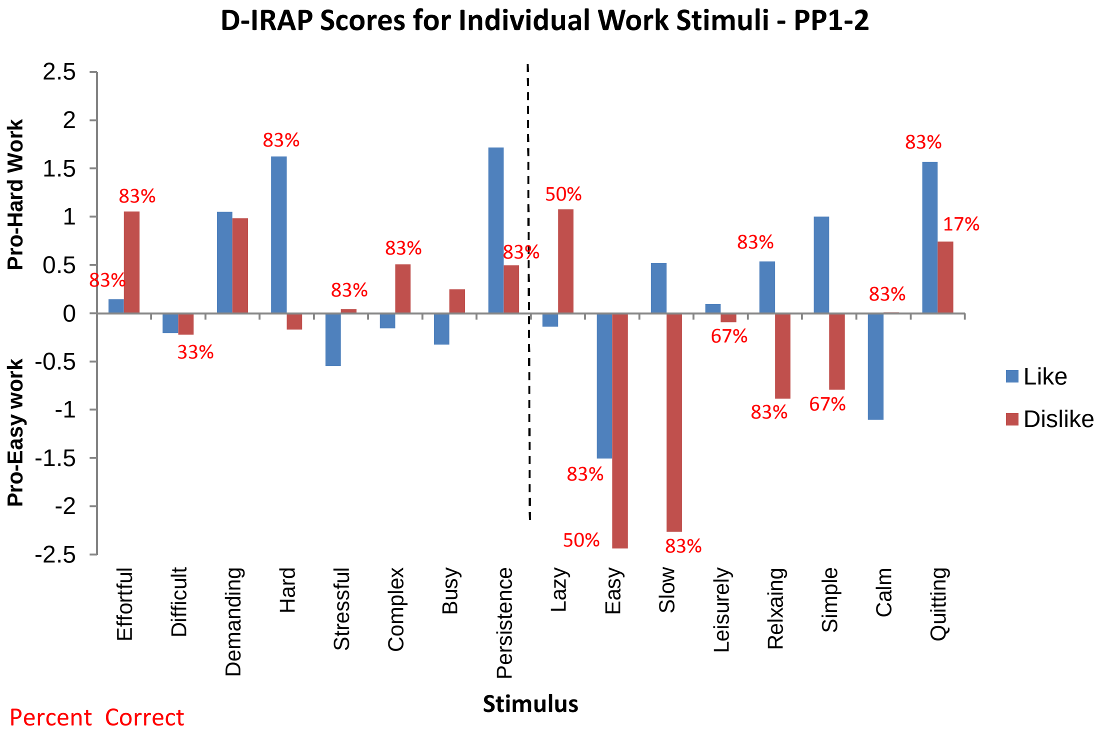


Figure 8. Stimulus-level D-IRAP scores for Participant PP1-9, representative of a positive correlation, for traditional IRAP (top panel) and MT-IRAP (bottom panel).

Percentage correct for stimulus combinations with errors shown in red for traditional IRAP; all other unmarked combinations had no errors (100% correct). Dashed line separates “hard work” from “easy work” words.


D-IRAP Scores for Individual Work Stimuli - PP1-2
Pro-Hard Work
Pro-Easy work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
83%
83%
33%
83%
83%
83%
83%
50%
83%
50%
83%
83%
67%
83%
67%
83%
83%
17%
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relxaing
Simple
Calm
Quitting
Stimulus
Like
Dislike
Percent Correct

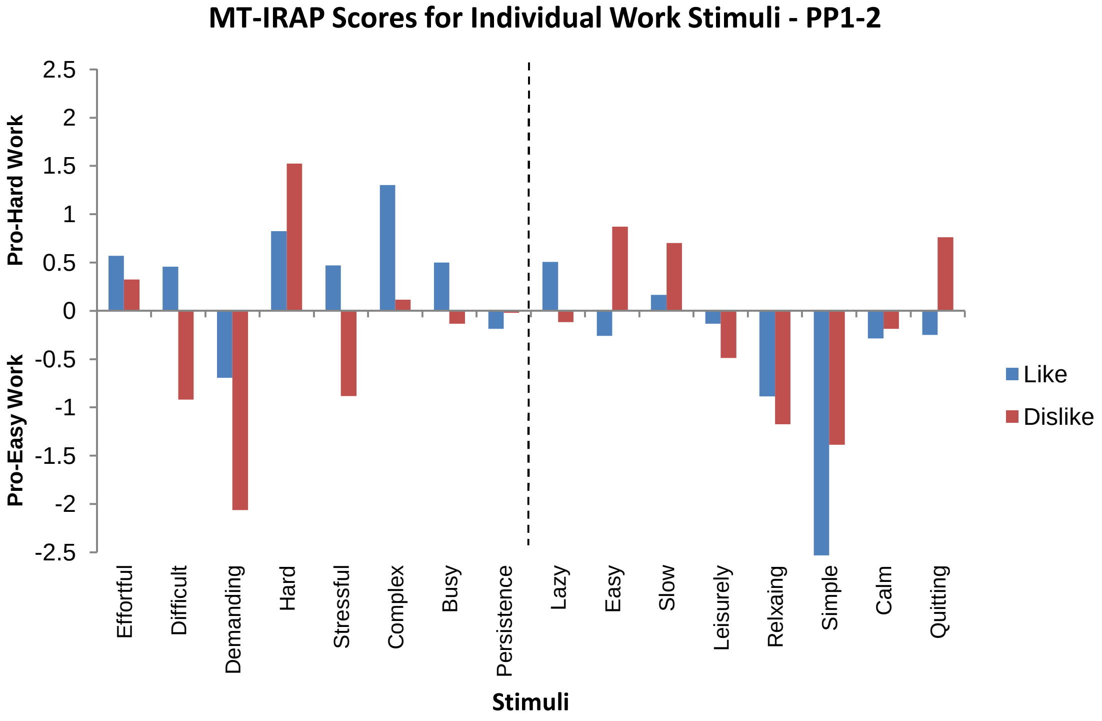
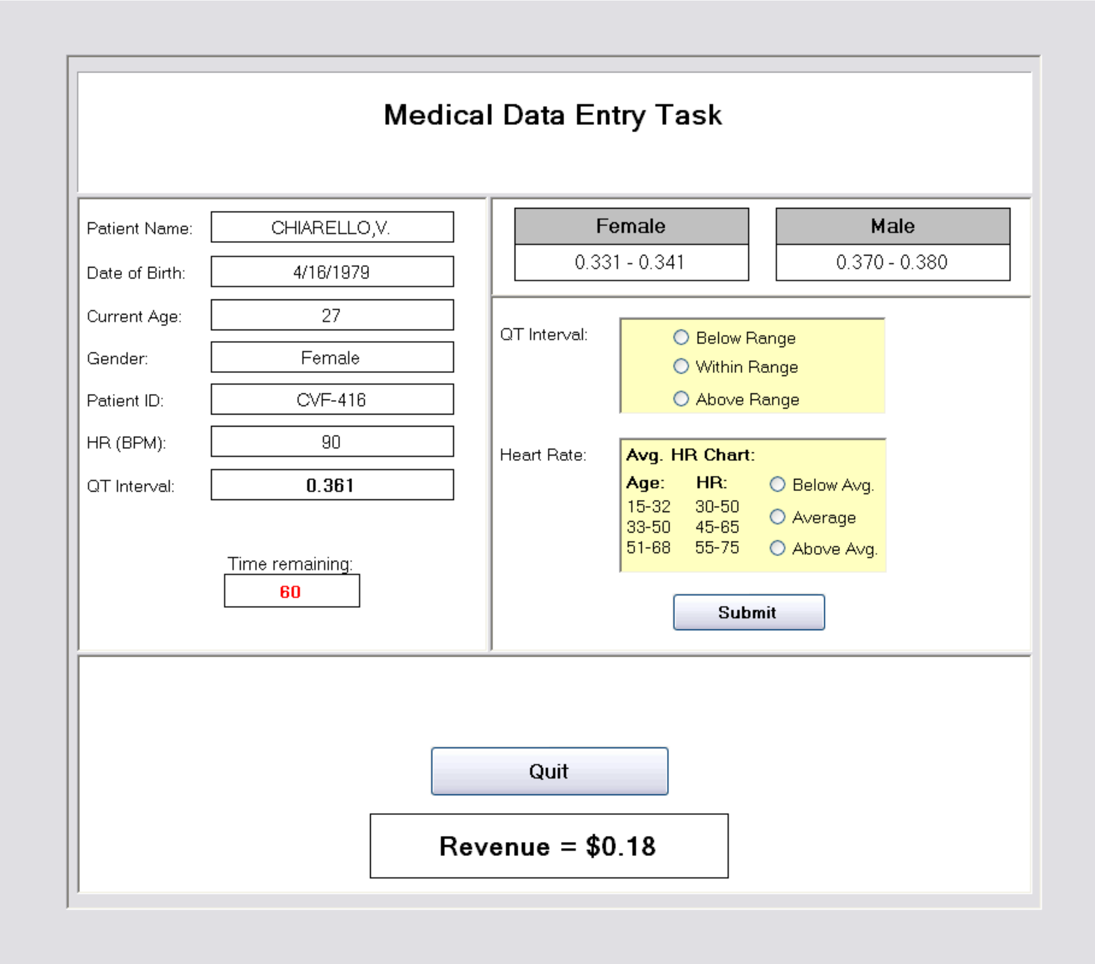

MT-IRAP Scores for Individual Work Stimuli - PP1-2
Pro-Hard Work
Pro-Easy Work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relxaing
Simple
Calm
Quitting
Stimuli
Like
Dislike

Figure 9. Stimulus-level D-IRAP scores for Participant PP1-2, representative of a negative correlation, for traditional IRAP (top panel) and MT-IRAP (bottom panel). Percentage correct for stimulus combinations with errors shown in red for traditional IRAP; all other unmarked combinations had no errors (100% correct). Dashed line separates "hard work" from "easy work" words.



Medical Data Entry Task

Patient Name:	CHIARELLO,V.
Date of Birth:	4/16/1979
Current Age:	27
Gender:	Female
Patient ID:	CVF-416
HR (BPM):	90
QT Interval:	**0.361**

Time remaining:
60

Female	Male
0.331 - 0.341	0.370 - 0.380

QT Interval:
○ Below Range
○ Within Range
○ Above Range

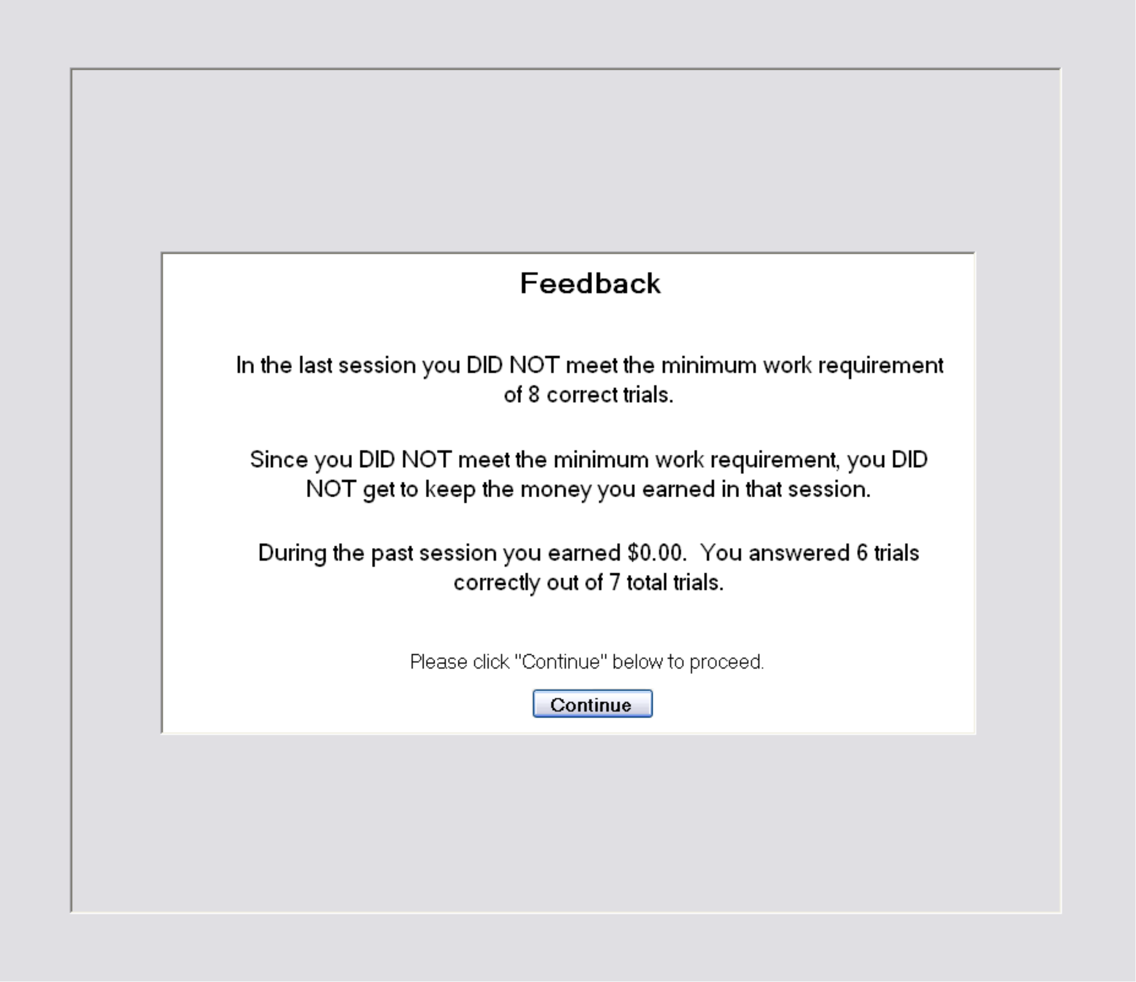
Heart Rate:
Avg. HR Chart:

Age:	**HR:**
15-32	30-50
33-50	45-65
51-68	55-75

○ Below Avg.
○ Average
○ Above Avg.

Revenue = $0.18

Figure 10. Screenshot of the simulated data entry work task.



Feedback

In the last session you DID NOT meet the minimum work requirement of 8 correct trials.

Since you DID NOT meet the minimum work requirement, you DID NOT get to keep the money you earned in that session.

During the past session you earned $0.00. You answered 6 trials correctly out of 7 total trials.

Please click "Continue" below to proceed.

Figure 11. Screen shot of feedback between 2-min work sessions in data entry work task.

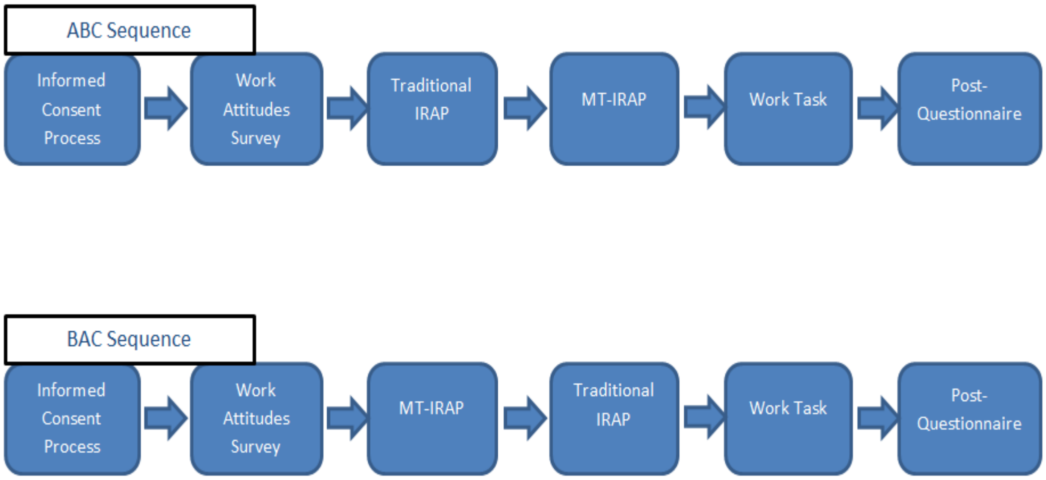
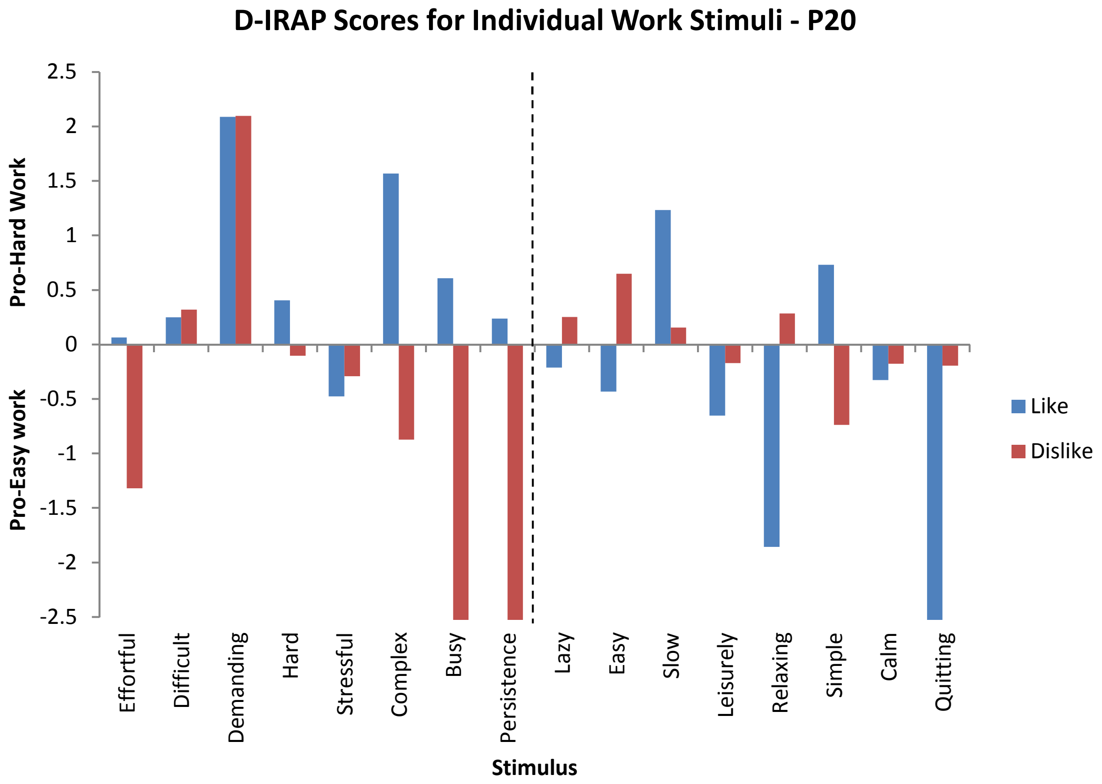


Figure 12. Visual depiction of experimental procedure for ABC and BAC sequences in Experiment III.


D-IRAP Scores for Individual Work Stimuli - P20
Pro-Hard Work
Pro-Easy work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimulus
Like
Dislike

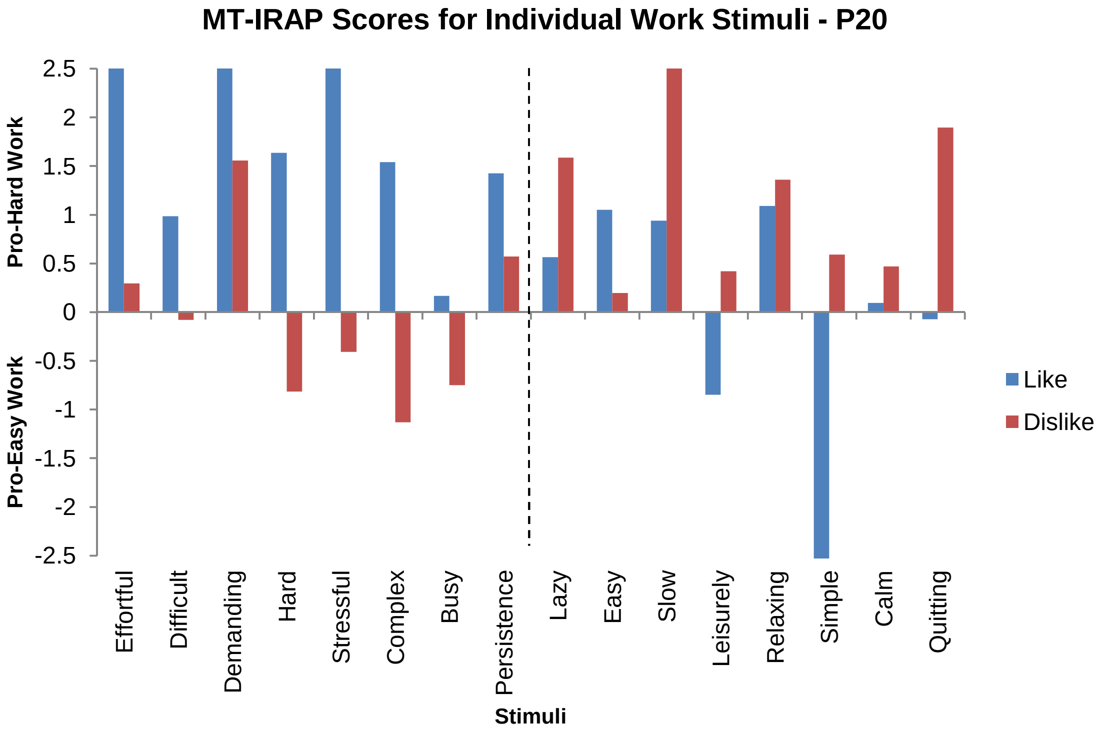
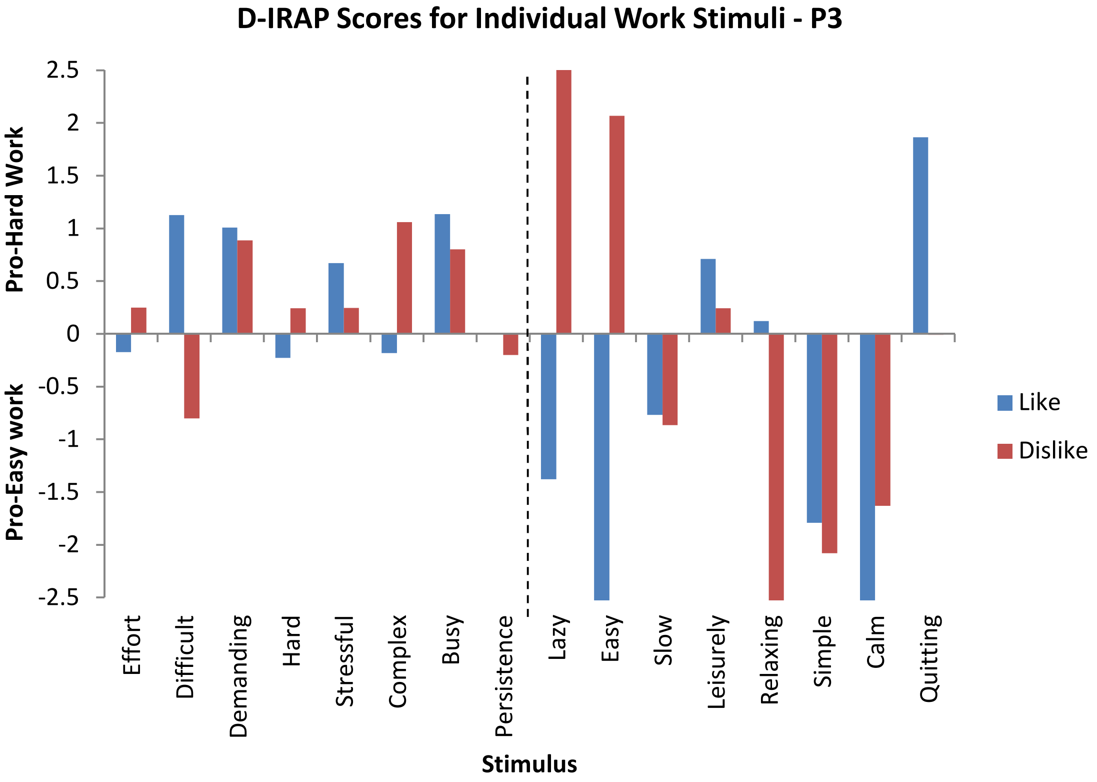

MT-IRAP Scores for Individual Work Stimuli - P20
Pro-Hard Work
Pro-Easy Work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimuli
Like
Dislike

Figure 13. D-IRAP scores at individual stimulus level for representative sample of positive correlation (*r* = .49; Participant 20) between traditional IRAP (top panel) and MT-IRAP (bottom panel) in Experiment III.


D-IRAP Scores for Individual Work Stimuli - P3
Pro-Hard Work
Pro-Easy work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effort
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimulus
Like
Dislike

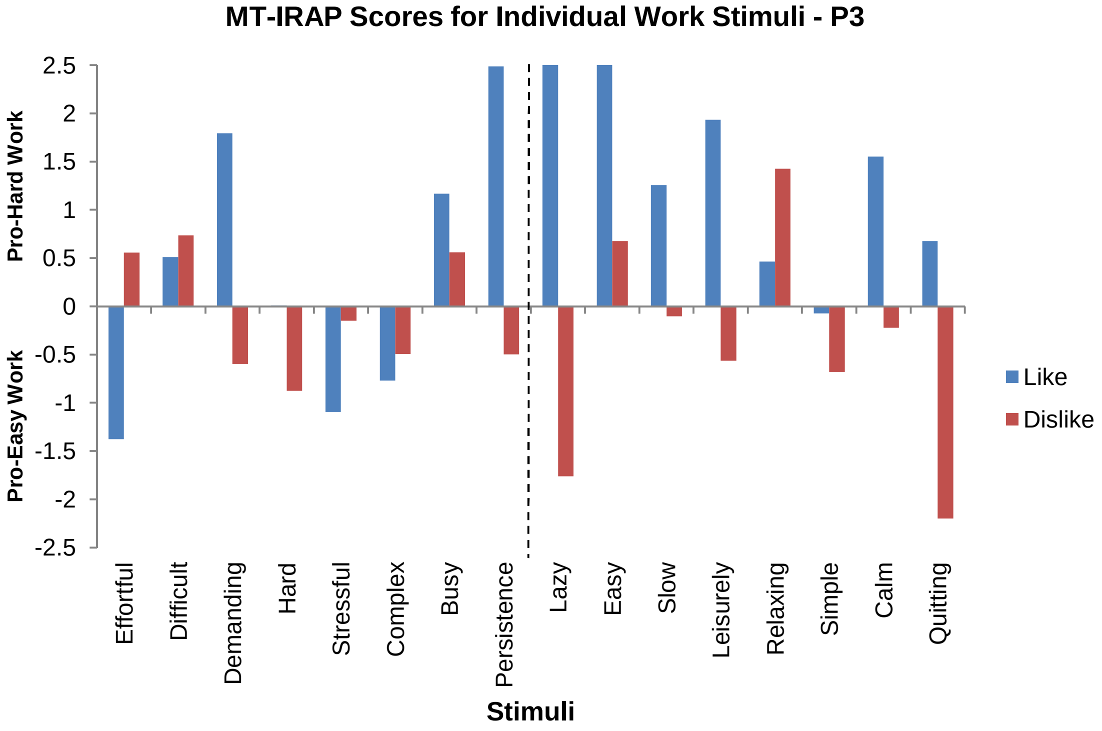
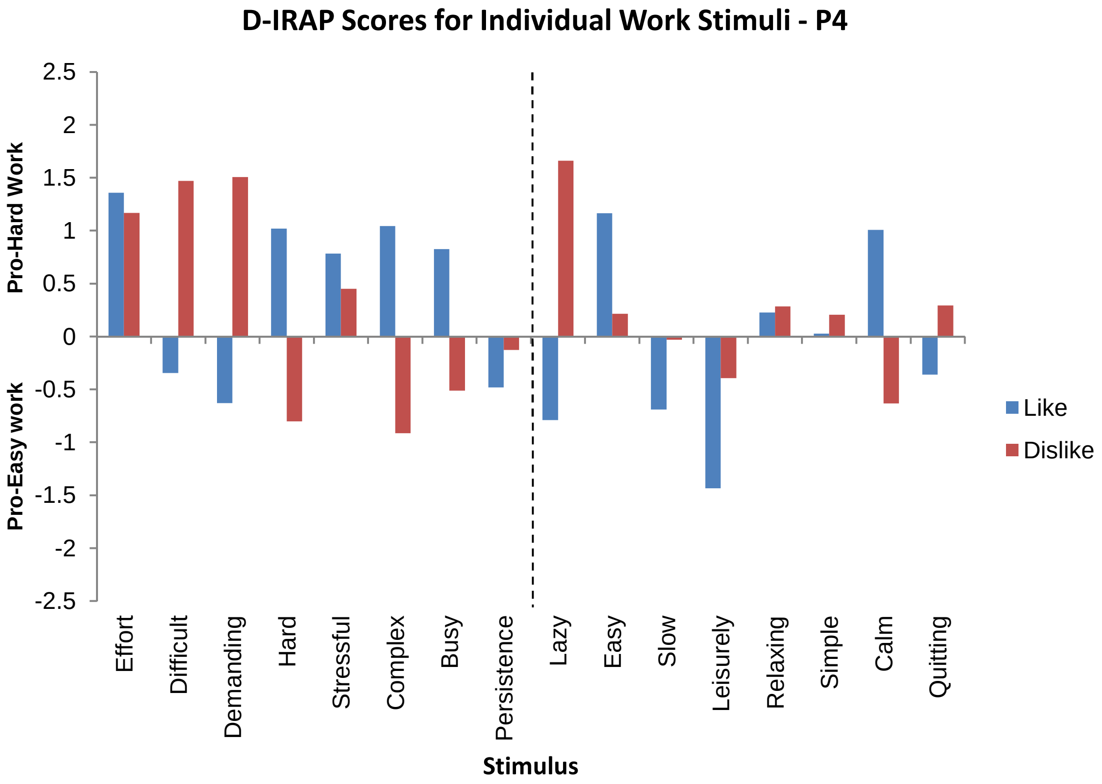

MT-IRAP Scores for Individual Work Stimuli - P3
Pro-Hard Work
Pro-Easy Work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimuli
Like
Dislike

Figure 14. D-IRAP scores at individual stimulus level for representative sample of negative correlation (*r* = -.38; Participant 3) between traditional IRAP (top panel) and MT-IRAP (bottom panel) in Experiment III.


D-IRAP Scores for Individual Work Stimuli - P4
Pro-Hard Work
Pro-Easy work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effort
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimulus
■ Like
■ Dislike

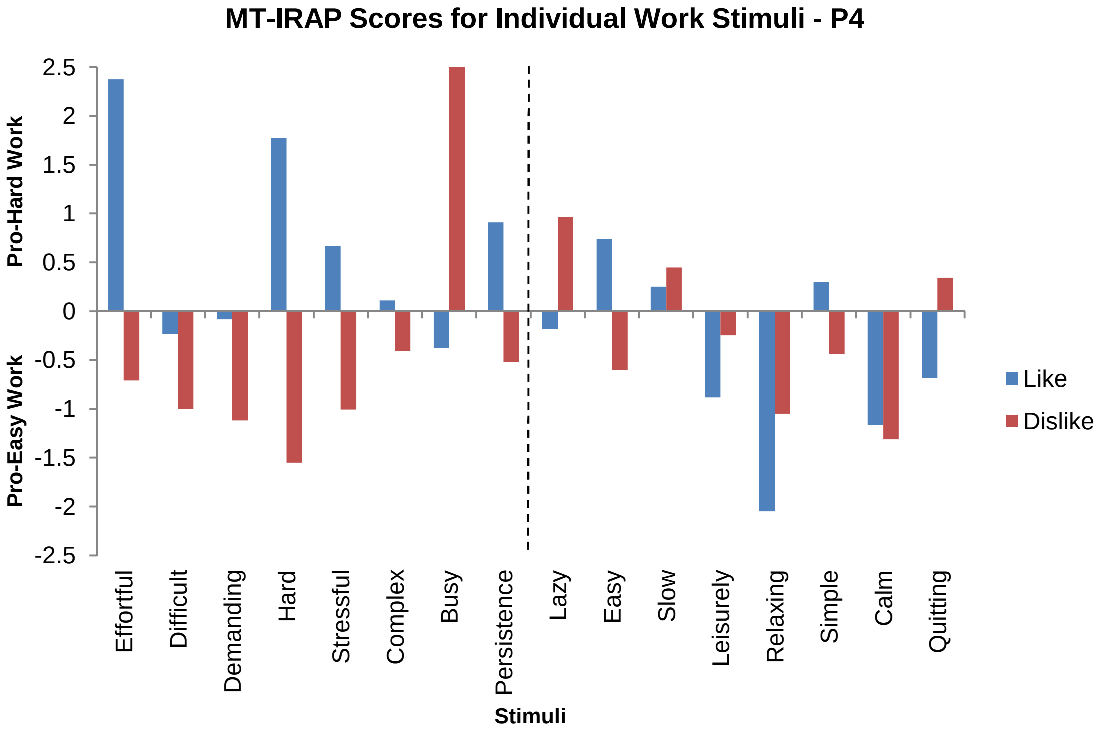

MT-IRAP Scores for Individual Work Stimuli - P4
Pro-Hard Work
Pro-Easy Work
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
Effortful
Difficult
Demanding
Hard
Stressful
Complex
Busy
Persistence
Lazy
Easy
Slow
Leisurely
Relaxing
Simple
Calm
Quitting
Stimuli
■ Like
■ Dislike

Figure 15. D-IRAP scores at individual stimulus level for representative sample of no correlation (*r* = -.01; Participant 4) between traditional IRAP (top panel) and MT-IRAP (bottom panel) in Experiment III.

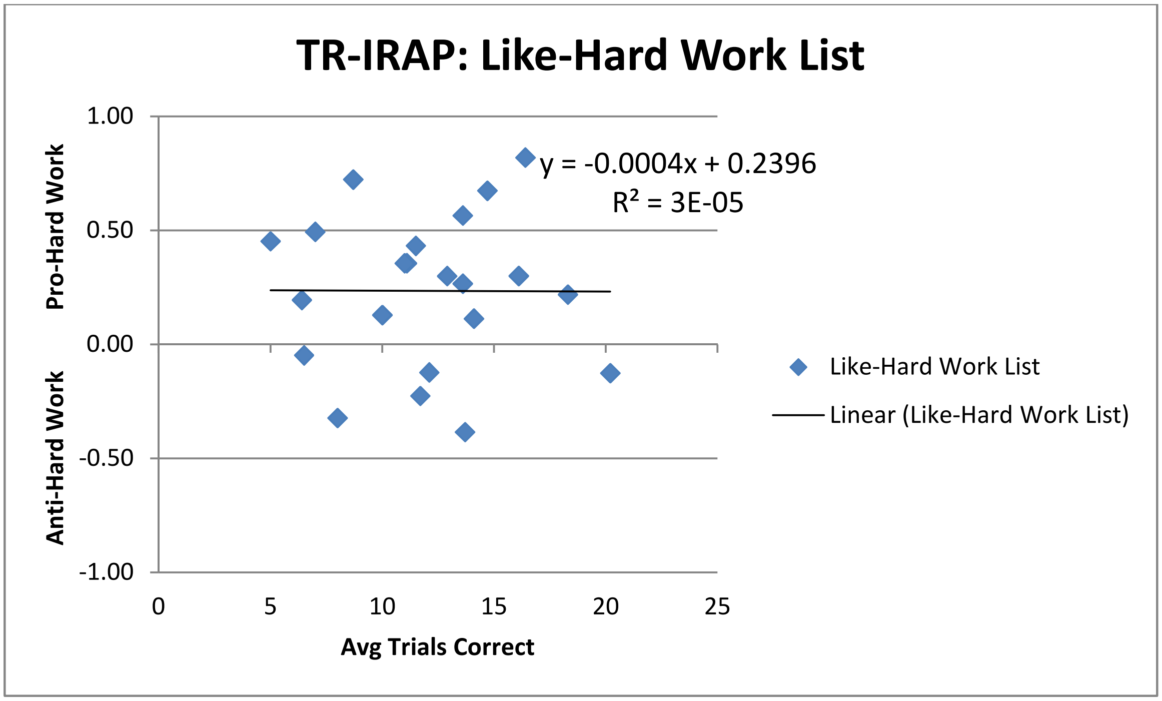


Figure 16. Scatter plot and regression function of traditional IRAP scores for Like-Hard Work list and average correct trials per session during work task in Experiment III.

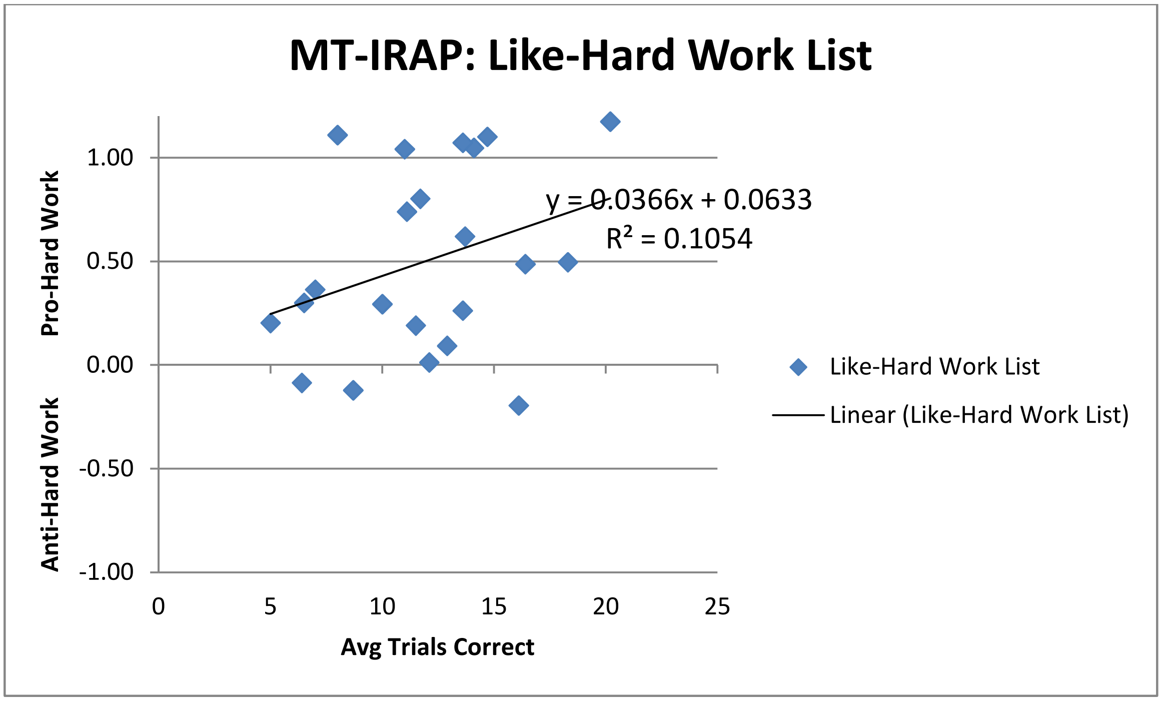


Figure 17. Scatter plot and regression function of MT-IRAP scores for Like-Hard Work list and average correct trials per session during work task in Experiment III.

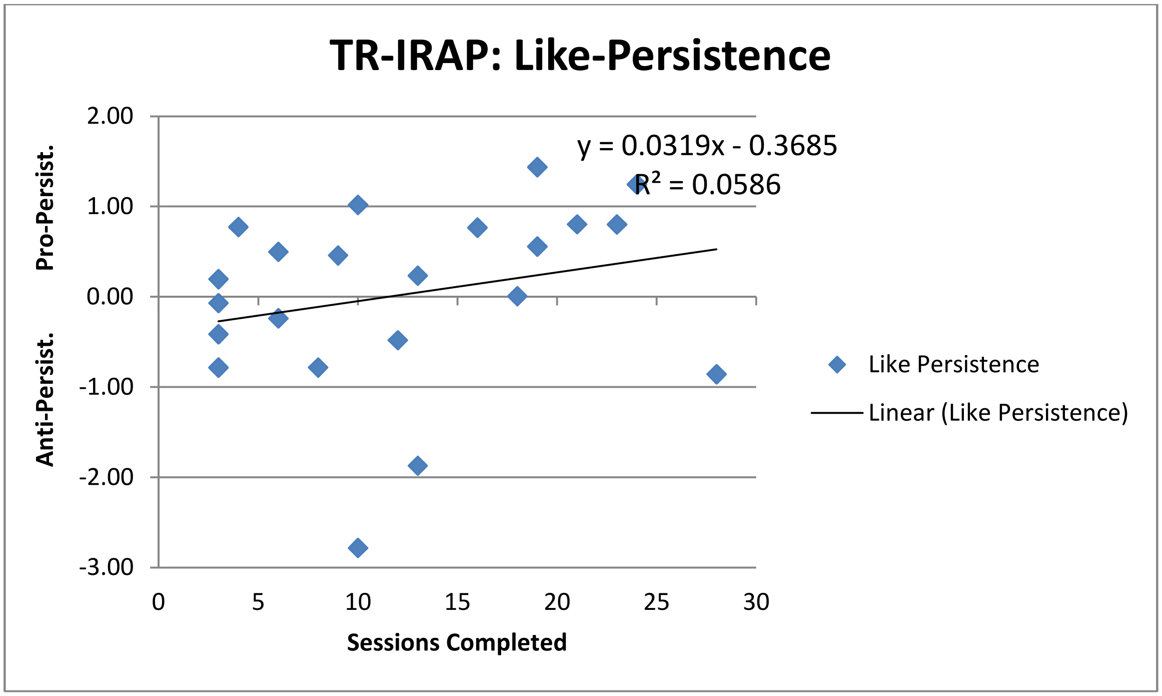


Figure 18. Scatter plot and regression function of traditional IRAP scores for Like-Persistence stimuli and number of sessions completed during work task in Experiment III.

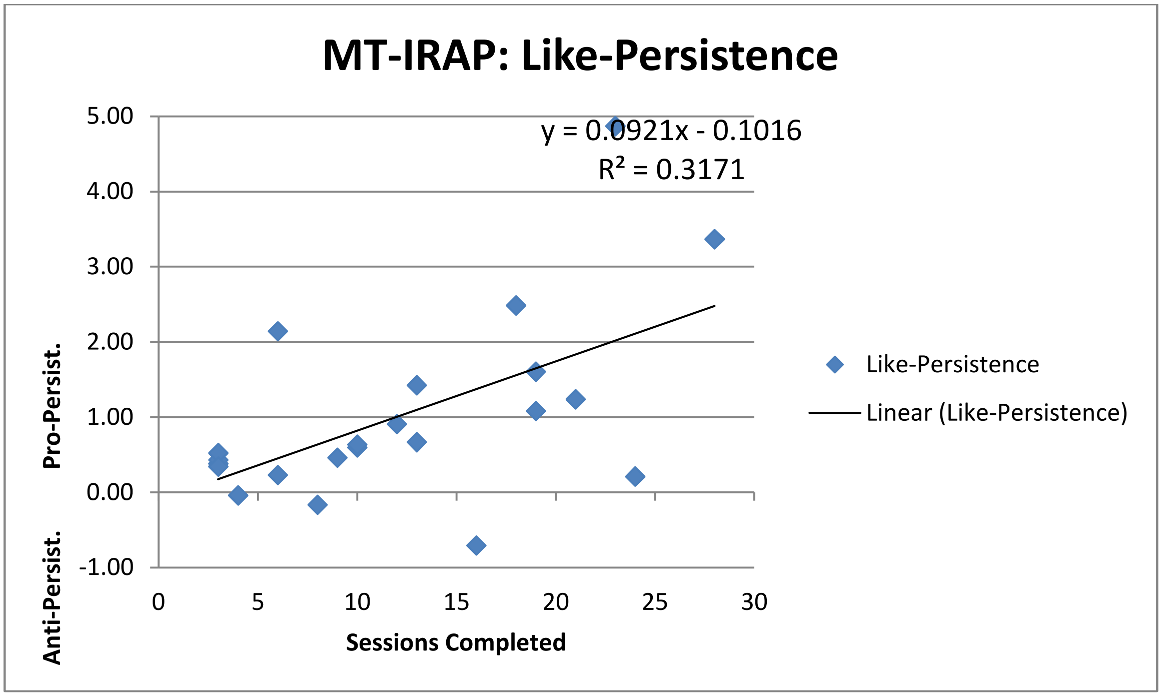


Figure 19. Scatter plot and regression function of MT-IRAP scores for Like-Persistence stimuli and number of sessions completed during work task in Experiment III.

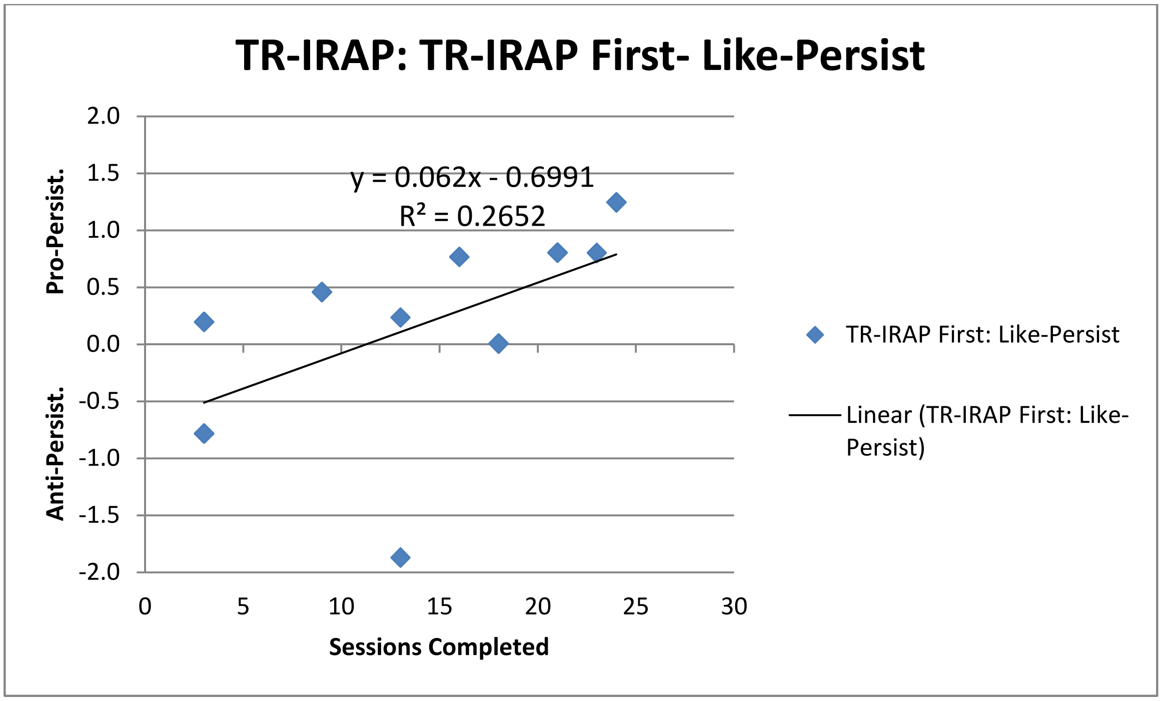


Figure 20. Scatter plot and regression function of traditional IRAP scores for Like-Persistence stimuli and number of sessions completed during work task for those participants who completed the traditional IRAP first (ABC sequence) in Experiment III.

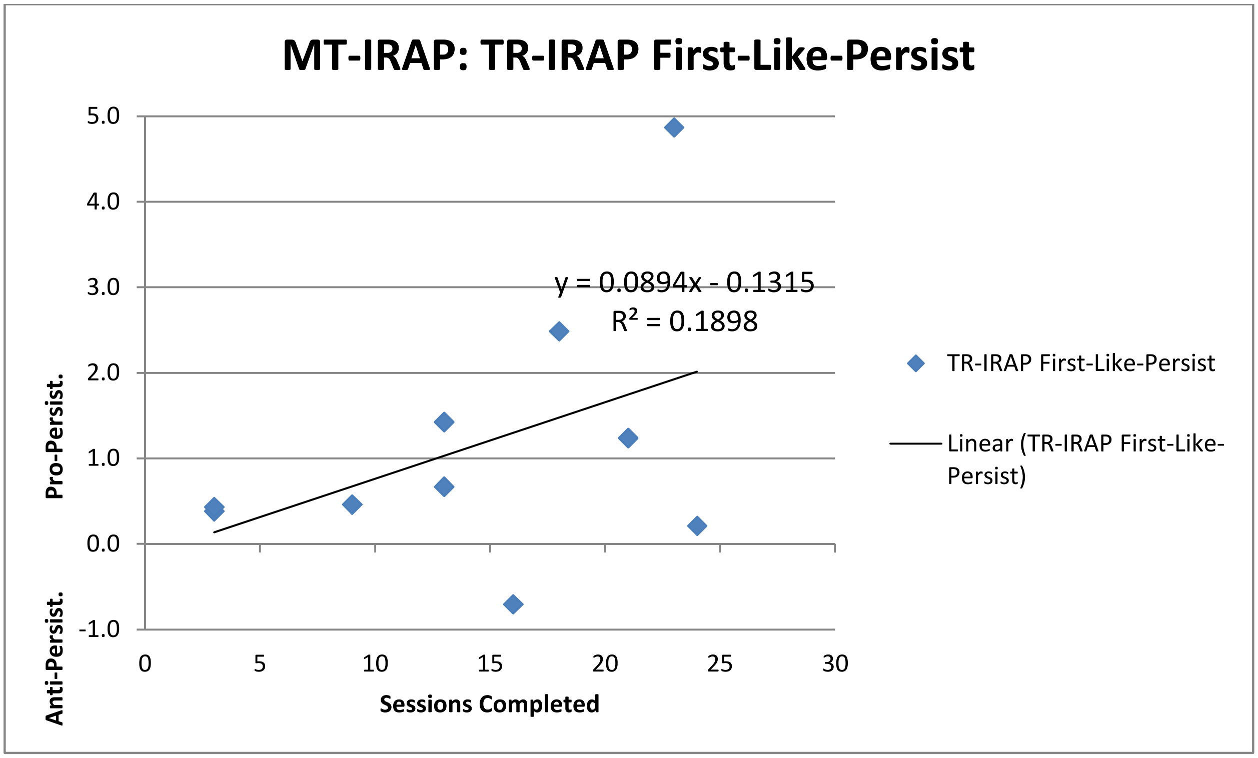


Figure 21. Scatter plot and regression function of MT-IRAP scores for Like-Persistence stimuli and number of sessions completed during work task for those participants who completed the traditional IRAP first (ABC sequence) in Experiment III.

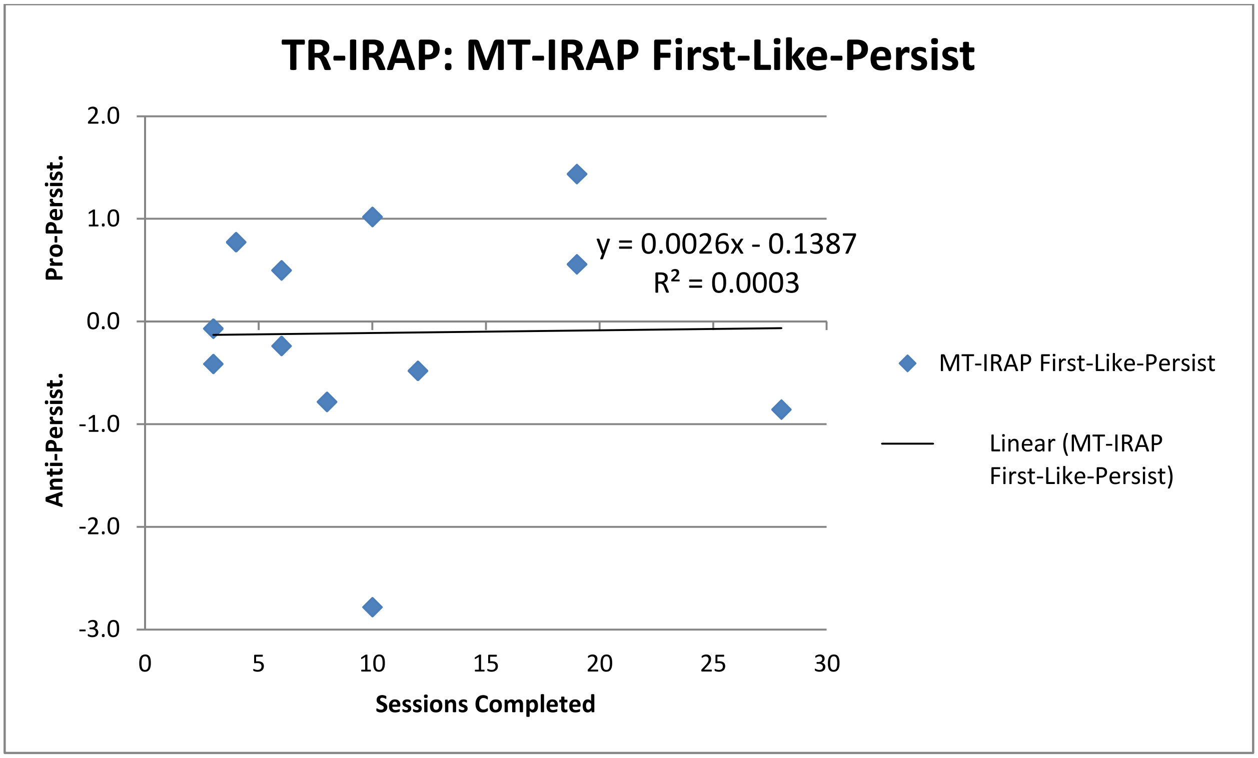


Figure 22. Scatter plot and regression function of traditional IRAP scores for Like-Persistence stimuli and number of sessions completed during work task for those participants who completed the MT-IRAP first (BAC sequence) in Experiment III.

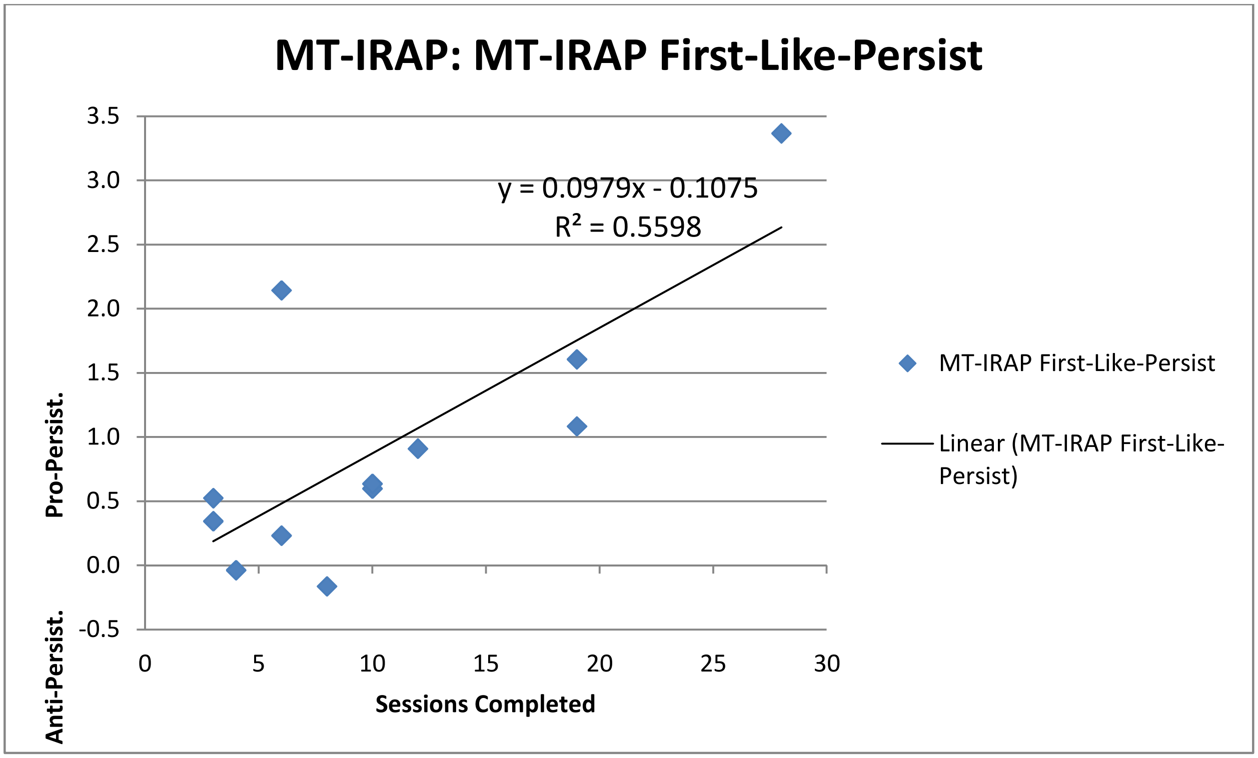


Figure 23. Scatter plot and regression function of MT-IRAP scores for Like-Persistence stimuli and number of sessions completed during work task for those participants who completed the MT-IRAP first (BAC sequence) in Experiment III.

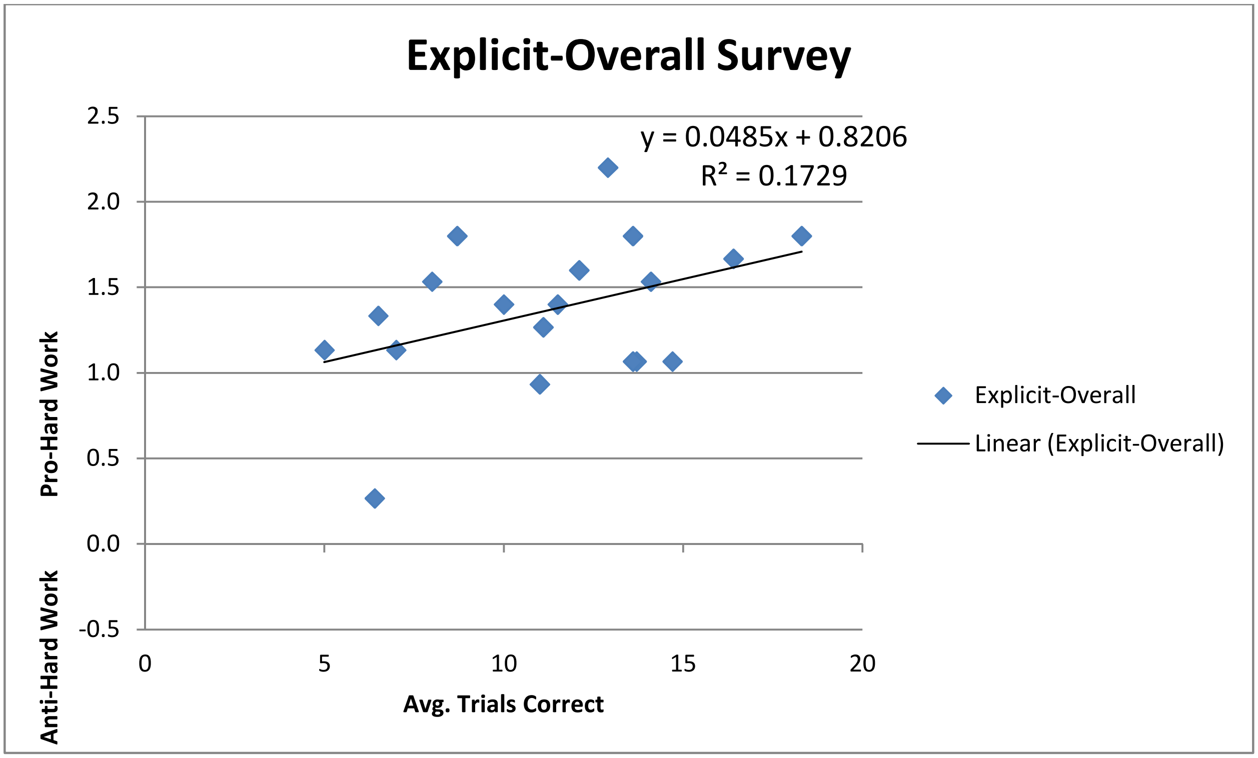


Figure 24. Scatter plot and regression function of overall explicit survey scores and average correct trials per session during work task in Experiment III.

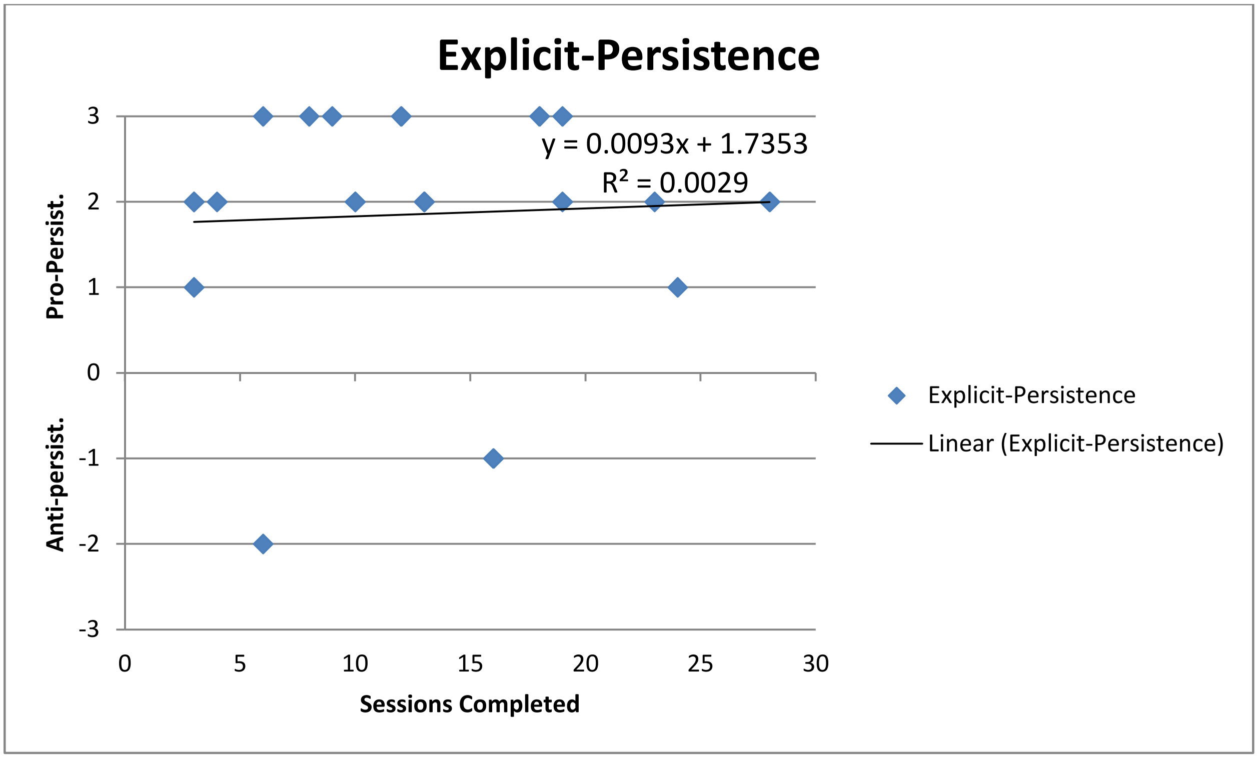


Figure 25. Scatter plot and regression function of explicit survey scores for Persistence stimulus and number of sessions completed during work task in Experiment III.

Appendix A

Instructions to Participants for Traditional IRAP

- **IMPORTANT:** Make sure you read EVERY instruction that appears on the screen, throughout the entire experiment. The instructions will tell you how to respond during each part of the experiment.
- During this experiment, you will see words appear at the top and middle portion of the screen. You will also see two different words ("Yes" and "No") appear at the bottom corners of the screen. Please ask the experimenter to show you a sample of the screen at this time.
- You will see words such as "Like" and "Busy" with the options "Yes" and "No" at the bottom. You should read these words as a sentence, such as, "I Like work that is Busy: Yes or No?"
- You must make the response of "Yes" or "No" as quickly as possible, using the 'D' and 'K' keys. **Note:** "Yes" and "No" will randomly switch back and forth between the 'D' and 'K' keys.
- If you respond too slowly, you will be told to go faster. If you make an incorrect response (pressing "Yes" when you should have pressed "No"), a red "X" will appear on the screen and you must press the correct response to move forward.
- As soon you finish one trial, the next one will begin immediately afterward. This experiment moves at a very fast pace and you must be ready for each trial immediately following the previous trial.
- Throughout the experiment, you will be told how to answer. Sometimes you will have to answer that you like words that have to do with hard work, and at other times you will have to answer that you like words that have to do with easy work. Instructions will appear on the screen periodically and will tell you how you should answer.
- You will first go through some practice trials to familiarize you with the task. If you pass practice, then you will begin the actual test portion of the experiment.
- Please ask the experimenter if you have any questions at this time. Get ready to begin!

Instructions to Participants for MT-IRAP

Pairs of words will be presented to you on the screen and you will be asked to match these words (e.g., "Like" and "Flower") to a particular response ("Yes" or "No"). You should read these words to yourself in a sentence, such as "I Like Flowers: Yes or No?" You will make your response of "Yes" or "No" using the 'D' and 'K' keys on the keyboard. The response associated with each key may change from one trial to the next (i.e., 'D' key could be "Yes" or "No" response on any given trial). You must make these responses each time as QUICKLY and ACCURATELY as possible. Be ready, as each trial will begin immediately following the last!

In addition to words like "Like" and "Flower," you will also see a third word in the center of the screen: either "Truth" or "Lie". When you see "Truth," you are to tell the TRUTH about how you feel, in terms of whether you "Like Flowers" or not. Similarly, when you see "Lie" in the center of the screen, you are to LIE about how you feel, so if you do "Like Flowers," then in this instance you should answer "No" (instead of "Yes"), since you are lying about how you really feel. At the beginning of every trial, before the other words appear on the screen, you will first see only "Truth" or "Lie," so that you know whether you are tell the truth or lie in the trial that is about to begin.

You will now go through some practice trials to familiarize you with the task. Remember, you must answer as QUICKLY and CONSISTENTLY as possible. If you answer too slowly, you will be told to answer more quickly. Please keep your fingers on the 'D' and 'K' keys at all times, so that you can respond as quickly as possible. The faster and more consistent you are in your answers, the sooner the assessment will be completed. If you are not fast or consistent enough, you may be asked to complete the assessment again.

This is Practice (I): Remember to respond to each set of words
as quickly as possible.

You will see words that pertain to flowers and insects. For the purpose of practice, it is assumed that you "Like" flowers and "Dislike" insects (e.g., spider). Therefore, in order to pass the practice section, you must respond that you like flower-type words and dislike insect-type words (even if you do not truly dislike insects, personally). For example, if you see "Dislike" at the top of the screen and "Cockroach" in the middle, along with the "Truth" label, then you must answer "Yes", indicating that you dislike cockroaches (this applies for the other insect words as well). During this practice section, you will only have "Truth" trials to answer.

In the event you answer incorrectly (for example, responding that you like insects), you will see a red "X" appear on the screen, and you must choose the other, correct response (Yes or No), before moving on. You must answer with an average accuracy of 70% and average response speed of ≤ 3 seconds, in order to successfully complete the practice sections. You will be allowed eight attempts to pass the two practice sections.

This is Practice (II): Remember to respond to each set of words
as quickly as possible.

During this practice section, you will have the same words (flowers and insects) with both "Truth" and "Lie" trials to answer. You must answer with an average accuracy of 70% and average response speed of ≤ 3 seconds, in order to successfully complete this practice section.

This is now the actual Test: Remember, you must respond to each set of words as quickly as possible.

You will see different words in the test section. For example, you will still see either "Like" or "Dislike" at the top of the screen, but instead of flower and insect words, you will see words pertaining to the workplace (e.g., "Hard Work"). You should read these words as a sentence, such as "I Like Hard Work: Yes or No?" During the test section, there is no predetermined right or wrong answer, so you should respond according to your own beliefs and attitudes. As such, no red "X" will appear after any of your responses in this section, however, you must still answer within 3 seconds or less.

In the test section, you will have both "Truth" and "Lie" trials to answer. You must answer as QUICKLY and CONSISTENTLY as possible. The more inconsistent your responding is with respect to your previous responses, the longer the assessment will take.

Appendix B

Work Attitudes Survey

Please answer the following questions as honestly as possible and indicate your answers by circling the answer that best represents your attitude.

1. At work, I tend to prefer tasks that are fairly easy.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

2. I don't like to exert myself or put too much effort into work.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

3. When I find that work seems too difficult, I don't like it.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

4. I am more likely to give up on work tasks that I find fairly difficult.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

5. Even in the work setting, I am someone who prefers to relax.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

6. I like work tasks that are fairly demanding.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

7. I tend to be lazy at work.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

8. I tend to persist at a task, even if it is difficult.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

9. I don't like being exposed to stressful work conditions.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly

Disagree		Disagree	Agree		Agree

10. I prefer pretty easy going work.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

11. I like working under calm work conditions.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

12. I don’t like to be pushed beyond what I’m capable of.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

13. I enjoy putting a strong effort into work.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

14. I like work that is complex and not too simple.

1	2	3	4	5	6
Strongly		Somewhat	Somewhat		Strongly
Disagree		Disagree	Agree		Agree

Appendix C

Post-Study Questionnaire

Please answer the following questions as honestly as possible. Please ask if you have any questions about what is being asked here.

1. Which was your primary motivation for signing up for this study?
   a. Extra credit
   b. Money
   c. Both
   d. Other: ____

2. Do you think the money you earned during the simulated work task influenced you performance on the task? Yes or No (circle one)

   a. Please explain: ____

   ____

   ____

   ____

3. What made you decide to quit the simulated work task? ______________________

   ____

   ____

   ____

4. Did you find the simulated work task to be difficult? Yes or No (circle one) Stressful? Yes or No (circle one) Did this change over time? Yes or No (circle one)

   a. Please explain: ____

   ____

   ____