Group analyses can hide heterogeneity effects when searching for a general model: Evidence based on a conflict monitoring task

Juan Botella[1], Jesús Privado[2], Manuel Suero[1], Roberto Colom[1]

and James F. Juola[1*]

[1]Universidad Autónoma de Madrid

[2]Universidad Complutense de Madrid

*** Corresponding author**

James F. Juola: juola@ku.edu

Abstract

In experimental psychology, a unique model of general processing is often sought to represent the behaviors of all individuals. We address the question of whether seeking this objective - a unique model - is the most fruitful scientific strategy by studying a specific case example. In order to approach an answer to such a question, we compared the conventional approach in experimental psychology with analyses at the individual level by applying a specific mathematical modeling approach. A sample of 1,159 individuals completed an experimental task based on managing conflict (a type of Simon task). Key findings revealed that at least four models are required to properly account for individuals' performance. Interestingly, four out of ten participants failed to show stimulus-response congruency effects in the experimental task, whereas the remaining 60% followed distinguishable theoretical models (consistent with conflict-monitoring theory and/or priming and episodic memory effects). The reported findings suggest that individuals' psychological characteristics might help to explain some of the reproducibility issues that are currently of great concern in psychology. These findings, along with further recent research, support the view that general and differential psychological approaches work better together for addressing relevant theoretical issues in psychological research.

Keywords: general psychology, differential psychology, cognition, modeling, Simon effect

1. INTRODUCTION

1.1. General and differential psychology

Attempts to combine general and differential approaches to the understanding of psychological phenomena has a long and venerable tradition (Cronbach, 1957, 1975; Sternberg, 1979). For instance, research has identified qualitatively different behaviors based on individual strategies or abilities in cognitive processes such as verbal comprehension (Hunt, 1978), visuospatial processing (Cooper, 1982), temporal order judgments (Grabot & van Wassenhove, 2017), deductive reasoning (Sternberg, 1980), analogical reasoning (Sternberg, 1977) and fluid reasoning (Carpenter et al., 1990). Such research has led to novel insights regarding different ways in which people cope with complex cognitive challenges; but perhaps more importantly, identification of qualitatively different subpopulations has allowed us to account for larger portions of the variance in overall performance (Nesselroade, 2010).

Assuming that there is only a single way to deal with a given cognitive task and, therefore, there exists a corresponding best model to describe behavior, might lead us to slippery conclusions. Inferences based on group data can sometimes be meaningless at the individual level (Estes, 1956). Research in experimental psychology and cognitive neuroscience is massively insensitive to this relevant possibility (Rapp, 2012), but the situation is changing quickly, at least in the latter discipline (Arend et al, 2003; Colom, 2016; Colom, Chuderski, & Santarnecchi, 2016; Dubois & Adolphs, 2016; Finn et al., 2015; Haier, 2017).

In a recent study, Grandy, Lindenberger, and Werkle-Bergner (2017) analyzed responses of 32 participants in the Sternberg (1966) short-term memory scanning paradigm. They first replicated the usual findings at the group level (linear increase in

overall mean reaction times as a function of memory set size), but afterwards they moved to the individual level, finding that the data from 13 participants were better fit by a self-terminating memory search model, whereas those from 13 other participants were better fit by an exhaustive memory search model. Therefore, in one group of just 32 individuals it was possible to identify instances supporting two of the most well-known competing models for describing the cognitive strategy employed in scanning short-term memory. The implication is straightforward: an adequate description of memory search is unlikely to be achieved with a single general model. It is even possible that the search for general models for tasks that are sensitive to competing strategies or competencies can lead to failures to replicate, in cases in which heterogeneous groups of subjects are mixed in different proportions across studies. At least part of the "replication crisis" in psychology could be due to a reluctance to seek explanations based on individual differences and heterogeneous behaviors (see, e.g., DeCoster, Sparks, Sparks, Sparks, & Sparks, 2015; OSC, 2015).

1.2. General models and individuals

Research in experimental psychology seeks general models for describing and explaining human behavior. These models manage relationships among variables in an attempt to identify the key factors to which the dependent variables might be sensitive. Experiments are usually designed to choose among competing models that make distinguishable predictions for a given phenomenon of interest (Danziger, 1990; Gigerenzer, 1987; Lamiell, 2003).

It is assumed by default (as a kind of implicit meta-postulate) that a single, best model for properly describing human behavior can be discovered. Nevertheless, there is at least one alternative assumption that deserves close attention. Perhaps the ideal of

reaching a single model that explains the behavior of all individuals is not yet within the reach of the theories we are developing; perhaps such a model cannot be justified, even in the long term. Parameters describing individual behaviors could change, meaning that individual differences would be reflected in how the various distinguishable factors act upon the dependent variables of interest (Voelkle, Brose, Schmiedek, & Lindenberger, 2014).

Thus, for instance, models of cognitive performance assume that increased demands on working memory capacity degrade perceptual, attentional, memory, and reasoning performance (Colom, Abad, Quiroga, Shih, & Flores-Mendoza, 2008; Martinez et al., 2011; Unsworth, Fukuda, Awh, & Vogel, 2014). However, it is known that the effects of these processing demands are far from homogeneous for different individuals, which raises reasonable doubts regarding the likelihood of finding a single, best-fitting theoretical model for everyone (Luck & Vogel, 2013).

Here we will present and discuss an example showing how the individual differences approach can enrich and complement the ways we come to understand cognitive phenomena that are highly relevant for experimental psychology. Based on previous findings and theoretical expectations, we predict that several models will be required for achieving proper descriptions of individuals' behaviors. There may be individuals sharing qualitatively different mechanisms for coping with the same cognitive challenge. Furthermore, there should also be quantitative differences within groups. Both results are likely and quite compatible.

1.3. Selective attention

Experimental psychologists have long debated where selective attention acts within the stream of cognitive processes. Broadbent's filter model (1958) set that locus

in the analysis of the physical properties of the stimuli. Deutsch and Deutsch (1963) set it after content or semantic analysis. Treisman (1964) proposed that the stimuli were not completely selected, but variably attenuated, and Lavie (2001) argued that the locus of the attentional filter could be flexible depending on cognitive load.

Nevertheless, it is reasonable to suspect that there could be individuals who perform exactly the same selective attention tasks following any of several possible models, or even some alternation or combination of them. All could be valid for describing cognitive performance at the individual level. The way different individuals approach the task might be hardwired or not (an issue we will discuss later on). We now simply emphasize that in testing general models against experimental data, different participants could complete tasks by following distinguishable paths that might lead us to favor a specific model while rejecting others that might well be more applicable to some other subset of the participants.

1.4. Predictions

Experimental data are often analyzed by computing linear models such as the analysis of variance (ANOVA). The hypotheses derived from the models are, however, only ordinal predictions for the population means. When rejecting the null hypothesis of no differences, H_0, an alternative, H_1, is supported which assumes specific ordinal relationships among the parameters. When this happens, it is concluded that the model from which those predictions are derived is (provisionally) 'correct'.

Now let's imagine that the model accepted is a good description for only some of the individuals. Still, the ANOVA could yield statistically significant differences that will be interpreted as supporting the model tested, even though the heterogeneity of the participants could go unnoticed. It is easy to predict that in future experiments with

small variations in the experimental paradigm, those individuals properly described by the previously accepted model will be less well represented in the new sample. The findings could well be puzzling and, instead of moving ahead, we will walk in circles.

This crucial problem has been already addressed, mainly by computations based on simulation studies in which the true model is known, since it is the model from which the data had been generated. Simulation studies allow assessing the consequences of analyzing the data in several alternative ways (Cohen, Sanborn & Shiffrin, 2008; Estes & Maddox, 2005; Lee & Webb, 2005; Smith & Batchelder, 2008, 2010).

Here we will describe by an example with real data how the problem can show up. We will also discuss how, when looking for a general model, group analyses can hide the presence of heterogeneity among the participants. Afterwards, we will raise some suggestions for coping with these troubling scenarios. We will rely on a large dataset of 1,159 individuals considered by Privado, Botella, Suero, Quiroga, and Colom (2015) in a task testing for sequential effects in a stimulus-response congruency experimental paradigm.

2. METHOD

2.1. Participants

Three successive samples totaling 1,159 were selected for participation from an undergraduate psychology class at the Universidad Complutense de Madrid (SA = 379, SB = 358 and SC = 422). Their mean age was 20.2 years (SD = 3.0, Range = 18 to 54). We kept the three samples separate for testing the replicability of the overall results.

2.2. Experimental task

Each individual completed 100 trials, as part of a larger series of experiments, with the first 20 considered as practice trials. In the chosen experimental paradigm, each participant was to respond, as quickly as possible while avoiding errors, to an arrow stimulus appearing to the left or to the right of a central plus sign serving as a fixation point on a pre-exposure field that remained on when the arrow appeared. The arrow could point to the left or to the right. The place where the target appeared was irrelevant for the task, and the participant was to respond with the corresponding left or right hand to the direction in which the arrow was pointing (Figure 1a). The displays in the upper left cell and the lower right cell of Figure 1a are congruent (congruency between the direction in which the sign is pointing and its position relative to the central plus sign), and the other two are incongruent (incongruence between the arrow's indicated direction and its relative location). Half of the trials were congruent and the remaining were incongruent trials. The sequence of 80 experimental trials was randomly set for each participant, which included 20 trials for each of the four displays. The experimental procedure is described in full detail in Privado et al., (2015).

----- Figure 1, about here -----

We expected to find a *congruency effect* (CE), in that the mean RT for the incongruent trials should be greater than that for the congruent trials (i.e., a Simon, 1969; 1990 effect). In addition, we were interested in the presence of any *sequential congruency effects* (SCEs), that is, whether the CE is moderated by the characteristics of the previous trial.

SCEs have been reported many times, with several different experimental paradigms and alternative explanatory models. Specifically, after an incongruent trial

the CE tends to be smaller than after a congruent trial (see Figure 1b). According to the *conflict monitoring theory* (Botvinick et al., 2001, 2004), congruent trials produce a relatively low level of conflict, and attentional control is temporarily relaxed. If the following trial is incongruent, competing tendencies for automatic stimulus-response associations compete with top-down control to produce interference and larger CEs. On the other hand, an incongruent trial produces a transient increase in top-down control, which translates into less influence of automatic stimulus-response associations on the next trial, producing a smaller CE.

Subsequent research has introduced a variety of other explanatory models for SCEs, including those that emphasize the episodic structure of the trial sequence. These successive episodes can induce sequential stimulus and/or response priming effects, or integration of stimulus-response features that persist to affect subsequent trials (e.g., Hommel, 2004, 2007; Hommel, Proctor & Vu, 2004; Mayr, Awh, & Laurey, 2003). Other models have emphasized the roles of contingency learning or expectations for repetition or alternations among the small numbers of trial types used, with their subsequent mediation of the observed CEs (see Duthoo, Abrahamse, Braem, Boehler, & Notebaert, 2014, for a review). The facts that a variety of models has been suggested, and each has gained some support in the literature, indicate that either the phenomenon to be explained is very complex, or there is a variety of strategies or abilities represented in the samples of participants studied. The latter possibility is the main focus of the present paper.

2.3. Analyses

2.3.1. Conventional analyses

We computed ANOVAs for testing whether there is a CE, and whether there is any SCE associated with the congruency condition of the previous trial or with any repetition of the previous stimulus. A main difference between conventional and individual analyses is that in the former the only values in the statistical analysis for a given condition are the mean response times (RTs) for each individual in that condition. When performing individual analyses (see the next section) we will use the values for each trial as data points.

We will refer to the RTs for congruent and incongruent trials with uppercase subscripts. For example, there is a CE (standard congruency effect) if,

$$\overline{RT}_I - \overline{RT}_C > 0$$

We represent the type of the previous trial with a lowercase subscript before the uppercase letter for the current trial type (for example, *cI* represents an incongruent trial preceded by a congruent trial). The expected SCE appears if the CE is larger after a congruent trial than after an incongruent trial; that is, a SCE is observed when

$$\left(\overline{RT}_{cI} - \overline{RT}_{cC}\right) - \left(\overline{RT}_{iI} - \overline{RT}_{iC}\right) > 0$$

For the conventional analysis (group analysis) we first calculated the mean RT for each participant in the four conditions of the experiment (after removing 7.5% of the trials as errors or outliers, the latter being defined as RTs > 2000 ms or < 200 ms). The performance of each individual in each of the conditions is represented by the mean RT for trials in that condition.

With those data we conducted repeated-measures, 2X2 ANOVAs for the four types of trials, according to two factors: the congruency of the current trial and the congruency of the previous trial. We have chosen this ANOVA model because it is the one most frequently found in studies published in recent decades. Currently, more realistic and powerful statistical tools, such as linear mixed models, are gaining popularity. However, we have preferred to use the statistical model that has been used most commonly from which evidence has been generated and evaluated as being contradictory and has led to the debate on replicability.

The priming hypothesis can be tested by removing the trials that are sequences of pure replicas (see Figure 1b). If the key for the SCE is in the exact repetition of the stimuli, then that effect should vanish when eliminating the pure replicas.

In the next set of analyses, we look for the effects at the individual level and compare the observed findings with those observed at the group level.

2.3.2. Individual analyses

A casual inspection of the individual data led us to think that the sample is composed of subsamples of individuals that complete the task in different ways. Some individuals performed as if the congruency of the previous trial was irrelevant, while for others the congruency of the previous trial was relevant, but not the physical match of the stimuli. It is also possible that some individuals fail to show any CE.

We next describe the procedures for classifying individuals in the various models that have been considered, taking as a basis the factors discussed above. The models assessed are designed to reflect alternative views of how different individuals can perform the same task. Each model is defined as a linear combination of factors that

express the variables already discussed (see Figure 2), and the parameters are estimated with linear regression methods.

----- Figure 2, about here -----

Model 0: no CE. The individual is insensitive to the congruency of the current trial and also to the congruency of the previous trial. Under this model individuals do not show any CEs. Expressing the mean RT of the individual under Model 0 in the long run as $\beta_{base(0)}$ and the random variations by the error component, *e*, the RT on trial *N* can be expressed as,

$$RT_{N(0)} = \beta_{base(0)} + e \quad [1]$$

Model 1: CE, but no SCE. The individual is sensitive to the congruency of the current trial, but not to the characteristics of the previous trial. Under this model individuals show significant CEs but no SCEs. The RT on trial N can be expressed as,

$$RT_{N(1)} = \beta_{base(1)} + \beta_{1(1)} \cdot X_{1(1)} + e \quad [2]$$

where $X_{1(1)}$ equals 1 if trial *N* is incongruent and 0 if it is congruent. The value of $\beta_{1(1)}$ reflects the size of the CE. The value of $\beta_{base(1)}$ in this model is not the general mean RT in the long run, but the mean RT in the long run for congruent trials. Model 0 is nested in Model 1, as the parameter in model 0 can be expressed as a function of the parameters in Model 1,

$$\beta_{base(0)} = \beta_{base(1)} + \frac{\beta_{1(1)}}{2}$$

Model 2: CE and SCE. The individual is sensitive to the congruency of the current trial and also to that of the previous trial. Furthermore, there is a possible interaction between the congruency conditions on both trials. This model follows the main

framework of the *conflict-monitoring theory* (Botvinick et al., 2001), since the CE is a function of the congruency of the previous trial, and the CE is expected to be smaller when the previous trial is incongruent than when it is congruent. The RT on trial N under model 2 can be expressed as,

$$RT_{N(2)} = \beta_{base(2)} + \beta_{1(2)} \cdot X_{1(2)} + \beta_{2(2)} \cdot X_{2(2)} + \beta_{3(2)} \cdot X_{3(2)} + e \qquad [3]$$

Where:

$X_{1(2)}$ equals 1 if trial *N – 1* is incongruent and trial *N* is incongruent; otherwise, 0.

$X_{2(2)}$ equals 1 if trial *N - 1* is incongruent and trial *N* is congruent; otherwise, 0.

$X_{3(2)}$ equals 1 if trial *N – 1* is congruent and trial *N* is incongruent; otherwise, 0.

$\beta_{1(2)}$ reflects the size of the SCE when trials *N - 1* and *N* are incongruent.

$\beta_{2(2)}$ reflects the size of the SCE when trials *N -1* is incongruent and *N* is congruent.

$\beta_{3(2)}$ reflects the size of the SCE when trials *N -1* is congruent and *N* is incongruent.

Note that $\beta_{1(2)}$ and $\beta_{3(2)}$ together reflect the size of the CE.

In this model $\beta_{base(2)}$ is the mean RT in the long run when both trials, *N* and *N - 1*, are congruent (the condition in which the shortest RT is expected). This model is more general than Model 1. Model 1 is nested in Model 2, as the two parameters in Model 1 can be expressed as a functions of the parameters in Model 2,

$$\beta_{base(1)} = \beta_{base(2)} + \frac{\beta_{2(2)}}{2}$$

$$\beta_{1(1)} = \frac{\beta_{1(2)} + \beta_{3(2)}}{2}$$

Model 3: CE and SCE based on stimulus repetition. The individual is sensitive to the congruency of the trial and to the repetitions of the stimuli in two consecutive trials. This model follows the main framework of the *priming hypothesis* (Hommel, 2004, 2007; Hommel, Proctor & Vu, 2004; Mayr, Awh, & Laurey, 2003). Pure replications

happen only if the two trials share the same congruency condition (the *cC* and *iI* sequences), but it does not happen in all sequences of those conditions. For example, (<+, <+) is a *cC* trial with repetition, whereas (+>, <+) is a *cC* trial without repetition. If the priming effect consists of reducing the mean RT in trials with pure replicas, then the CE will be smaller after an incongruent than after a congruent trial (SCE). This happens because of a reduction in the RT produced in about half of the trials of the *cC* and *iI* conditions (the pure replicas). The RT on trial N can be expressed under Model 3 as,

$$RT_{N(3)} = \beta_{base(3)} + \beta_{1(3)} \cdot X_{1(3)} + \beta_{2(3)} \cdot X_{2(3)} + e \qquad [4]$$

Where:

$X_{1(3)}$ equals 1 if trial *N* is incongruent and *N – 1* and *N* trials have the same stimuli; otherwise, 0.

The dummy variable $X_{2(3)}$ equals 1 if *N – 1* an *N* trials are different in congruency, otherwise, 0.

The parameter $\beta_{base(3)}$ is the mean RT in the long run for congruent trials when the stimuli in trials *N* and *N - 1* are exactly the same. The parameter $\beta_{1(3)}$ reflects the size of the effect when trial *N* is incongruent. The parameter $\beta_{2(3)}$ reflects the size of the effect when the stimuli are not repeated.

Model 1 is nested in Model 3, as the parameters in Model 1 can be expressed as functions of the parameters in Model 3,

$$\beta_{base(1)} = \frac{\beta_{base(3)} + (\beta_{base(3)} + 0.5 \cdot \beta_{2(3)})}{2}$$

$$\beta_{1(1)} = \beta_{1(3)} + \frac{\beta_{2(3)}}{2}$$

Model 4: CE and SCE based on congruency and repetition. This is the most general model; all the other models are nested in it. It considers that the individual is sensitive to

the interaction between the stimulus repetitions, the congruency condition on trial *N*, and the congruency of trial *N*-1. The RT on trial *N* can be expressed as,

$$RT_{N(4)} = \beta_{base(4)} + \beta_{1(4)} \cdot X_{1(4)} + \beta_{2(4)} \cdot X_{2(4)} + \beta_{3(4)} \cdot X_{3(4)} + \beta_{4(4)} \cdot X_{4(4)} + \beta_{5(4)} \cdot X_{5(4)} + e \quad [5]$$

Where:

$X_{1(4)}$ equals 1 if *N - 1* and *N* are congruent and not identical; otherwise, 0.

$X_{2(4)}$ equals 1 if *N - 1* and *N* are incongruent and identical; otherwise, 0.

$X_{3(4)}$ equals 1 if *N - 1* and *N* are incongruent and not identical; otherwise, 0.

$X_{4(4)}$ equals 1 if *N - 1* is incongruent and *N* is congruent; otherwise, 0.

$X_{5(4)}$ equals 1 if *N - 1* is congruent and *N* is incongruent; otherwise, 0.

Here $\beta_{base(4)}$ is the mean RT in the long run when the *N* - 1 and *N* trials are both congruent and the stimuli are repeated. The parameters $\beta_{\bullet(4)}$, where $\bullet$ takes values from 1 to 5, reflects the size of the respective effects.

All previous models are nested in Model 4, as all the parameters in those models can be expressed as linear combinations of the parameters included in Model 4. The equivalences between the parameters in Models 2 and 4 are,

$$\beta_{base(2)} = \frac{\beta_{base(4)} + (\beta_{base(4)} + \beta_{1(4)})}{2}$$

$$\beta_{2(2)} = \beta_{4(4)}$$

$$\beta_{1(2)} = \frac{\beta_{2(4)} + \beta_{3(4)}}{2}$$

$$\beta_{3(2)} = \beta_{4(4)}$$

The equivalences between the parameters in Models 3 and 4 are,

$$\beta_{base(3)} = \beta_{base(4)}$$

$$\beta_{1(3)} = \beta_{2(4)}$$

$$\beta_{2(3)} = \frac{\beta_{1(4)} + \beta_{3(4)} + \beta_{4(4)} + \beta_{5(4)}}{4}$$

We applied equations [1] through [5] to fit linear regression models to the RT data of each individual. Before fitting these least-squares, linear regression models, we checked for self-correlation in the sequence of data for each individual. The lag 1 autocorrelations were very low (the averages were .09, .09 and .10 for the three samples; the percentage of variance accounted for by self-correlation was less than 1% on average). Only 15% of the individuals showed a significant autocorrelation, and many of these were most likely to be Type I errors. Thus, from this point we ignore any problem related with self-correlation.

The selection of a best-fitting model for a given individual was accomplished by comparing their R^2 values, taking into account the residual squared sums and the number of parameters of each model. We have employed the following formula (Maxwell & Delaney, 1990), to choose between each pair of models,

$$\frac{\left(\sum e_{i_R}^2 - \sum e_{i_A}^2\right) \Big/ (df_R - df_A)}{\sum e_{i_R}^2 \Big/ df_A} \qquad [6]$$

The values obtained with [6] are distributed as $F(df_A\text{-}df_R, df_A)$, where df_A is the degrees of freedom of the model with more parameters (or augmented) and df_R is the degrees of freedom of the model with fewer parameters (or reduced). The degrees of freedom of a model is M-n, where M is the number of trials and n is the number of independent parameters. The number of trials was the same for each model, whereas the number of parameters is: 1 for Model 0, 2 for Model 1, 4 for Model 2, 3 for Model 3, and 6 for Model 4.

The rules for classifying a given individual are as follows (Figure 2). All participants were initially considered by default to be described by Model 0, the simplest one. Those for which the reduction in errors was significant with formula [6] for Model 1, were moved to that model. The same was done for Models 2 and 3, as related to Model 1. If in both cases the reduction in errors (increment of R^2) was not significant, the individual remained in Model 1. If it was significant for Model 2 and/or Model 3, the individual was classified accordingly. Sometimes an individual showed better fits for both Models 2 and 3. Finally, when Model 4 increased R^2 significantly when related to Models 2 or 3, the individual was classified in Model 4.

3. RESULTS

3.1. Results from conventional analyses

The ANOVA results are shown in Table 1. We show only results relevant for the following discussion: the main effect of congruency and the interaction between congruencies of the current and previous trials.

----- Table 1, about here -----

Both effects, CE and SCE, are present in all three samples. The size of the CE is 51 ms for the combined sample. The direction of the interaction is obvious in Figure 3a, which shows the data for the combined samples when replications are included: the CE is larger after congruent trials (94 ms) than after incongruent trials (10 ms); thus, the SCE equals 84 ms.

----- Figure 3, about here -----

3.2. Priming hypothesis

After computing the analyses regarding the priming hypothesis, we found the same significant effects, as shown in Table 2.

----- Table 2, about here -----

When eliminating the pure replicas, the CE is 54 ms for the whole sample, and the direction of the interaction is the same as before: the CE is larger after a congruent trial (79 ms) than after an incongruent trial (29 ms). Thus, the SCE equals 50 ms. Notice that the SCE is smaller when pure replicas are removed, but it is still significant. Figure 3b shows the effects after eliminating the trials with pure replicas.

The fact that the effect size decreased when pure replicas were eliminated, but the same effects remained significant, led us to the following conclusions for the three samples: (a) there is a significant CE in this spatial congruency task; (b) there is also a significant SCE associated with the congruency condition of the previous trial; and (c) when the displays of two consecutive trials are identical, the SCE is stronger than when the congruency condition is the same but the stimuli are different. From this type of analysis general conclusions can be reached about of the "average individual".

3.3. Individual level

Results derived from classifications detailed in the analysis section are shown in Table 3 (for the three samples) and Figure 4 (for the combined group).

----- Table 3 and Figure 4, about here -----

The distributions found after classifications in the three samples are very similar; the second and third samples replicated the results of the first sample. Perhaps the most surprising result is that the data from about four out of ten participants fit Model 0 best. That is, they failed to show any significant CEs; there were even 69 individuals who actually showed, at a descriptive level, larger average RTs on congruent than on incongruent trials; however, none of the congruency effects were significantly different from 0.0 for participants whose data were fit best by Model 0. Note that we have not made any correction to avoid an eventual increase in false positives when performing individual tests. Any correction of this type would have entailed more conservative decisions and, therefore, the rate of individuals not showing a significant effect of compatibility would have been greater.

Furthermore, (a) almost three out of ten participants were classified into Model 1, showing congruency effects (CEs) but no sequential congruency effects (SCEs), and (b) three out of ten participants show SCEs (including those sensitive to the repetition of stimulus congruency, those showing an effect of the repetition of the stimuli employed, and individuals showing both effects).

4. DISCUSSION

4.1. One model fails to fit all individuals

Here we have shown that the assumption of a single general model for describing how individuals process information for coping with a specific cognitively-demanding task can be misleading. Group analyses based on the presumption that all of them are essentially alike can easily lead to questionable conclusions. Under this framework, only quantitative individual differences reflected by the size of the effect or the impact

of some moderating factors are admissible. This strategy for analyzing the evidence can sometimes lead to a frustrating inferential indeterminacy (Smith & Little, 2018).

In the example considered here, group or conventional analyses led to the conclusion that participants showed congruency effects (CE), that they also showed sequential congruency effects (SCE) associated with the repetition of the condition of congruency, and that the SCE remained - but was smaller - when trials with stimulus repetitions were removed. Therefore, group analysis leads to the conclusion that, generally speaking, 'the average individual' shows all of these effects. A good model of human cognition when performing this task should include elements dealing with all of them as well.

However, when analyzed at the individual level, the results revealed a remarkably different picture. Indeed, a large number of participants failed to show any significant congruency effects (CEs). The CE has been reported in a variety of experimental paradigms (flanker tasks, Stroop effects, the Simon effect, and global/local processing, to name a few) in literally hundreds of experiments. Nevertheless, we must address the legitimate question of why such a large number of individuals do not show CEs, and also if this is a stable characteristic of individuals across congruency paradigms.

We also found that among the 60% showing the expected CE, only 50% showed a significant SCE. Again, this requires some explanation. What is the nature of the difference between the cognitive processes behind the responses of individuals showing or failing to show SCEs? We also must explore whether the SCE is absent across congruency paradigms in the same individuals. The same research logic applies to those individuals whose responses were sensitive to the repetition of the congruency condition and/or to the exact repetition of the stimuli employed in the current and previous trials.

4.2. The question of 'why'

Why are some individuals' behaviors fit properly by one given model, while others fare better if their behavior is described by some alternative model? This question is of paramount relevance, beyond Type I and Type II errors in classification. At the very least, we see two tentative answers.

First, we assume that there are differential choices regarding the strategy for coping with the same cognitive requirements. If this is likely, the 'true model' would encompass the models tested as representative of particular cases of the 'general model' (Figure 2).

Testing the likelihood of this first possibility requires showing that we can change the strategy spontaneously chosen by the individuals by modifying the characteristics of the task, the consequences following different types of errors, the instructions given, and so forth. Nevertheless, while developing such a general model we can still walk in circles when testing particular cases of the model (probably by considering them as rival models) and finding puzzling results when manipulating specific factors.

The second possibility is that differences between subsamples reflect enduring, perhaps hardwired characteristics. This perspective is similar to comparisons across people with different blood types (O, A, B, AB). The characteristics of each group make them eligible as donors for specific groups of recipients. They define 'qualitatively different groups', although there also might be quantitative differences within each group in some important characteristics (Lee & Webb, 2005).

Moving from daring analogies with blood to cognition, let's assume that there are groups of individuals with enduring qualitative differences, and that these differences are reflected in their behavior when completing tasks such as those regularly administered in experimental psychology. If this is the case, then results obtained with

conventional group analyses would be simply a function of the proportions of individuals sampled within each type in the experiment considered by any given study. Changes in these proportions will lead to difficulties in replicating results, not because we did something wrong, but because people are different.

However, the previous distinction between temporary or strategic differences and enduring differences is not essential for our argument here. Whatever their origin and nature, the fact is that we can be faced with samples of individual behaviors that cannot be accommodated by a single model. If we are not aware of this heterogeneity and implicitly assume the meta-postulate that there is a general model capable of explaining the behavior of all individuals, we can overlook crucial explanatory aspects at the theoretical level.

4.3. Consequences for reproducibility

One of the consequences of the failure to take into account individual and group evidence simultaneously is a potential impairment of the reproducibility of the phenomena under study. Attempts to reproduce the phenomena can lead to disappointing inconsistencies due to sampling different processing modes present in different proportions in the population of individuals. Suppose there exists a mixed population in which 60% of individuals' behaviors follow a certain model (M1) under which an explanatory factor has an effect that can be expressed as an effect size of d = 0.50; on the contrary, the behavior of the other 40% is explained by another model (M2) in which this factor is irrelevant, d = 0. When large samples of participants are used, the proportions of the two subpopulations in the sample will not be far from 60/40, but often the samples are not so large. To reach a power of 0.75-0.80 in an experiment with a within-subjects design with a medium effect size (d = 0.50), the sample should be

about 25 participants. In a sample of this size, the probability that the smaller subpopulation accounts for at least 50% of the sample is equal to 0.154. That is the same probability that the size of the average parametric effect of a random sample is less than 0.25. The expected value of the mean effect size in random samples is equal to 0.30. In short, the mere sampling of individuals from this mixed population adds a level of heterogeneity in the results that can be highly confusing. Researchers can waste their time looking for circumstantial features of the experiments with which to explain the differences. But, at least in some cases, the puzzle can be solved simply by questioning the meta-postulate that a single model must explain the behavior of all individuals. In order to converge on a single general model, it is essential that the conclusions derived from the analysis of the individual patterns of behavior be consistent with those of the pattern reflected in the sample means.

4.4. Conclusion

The present report leads to the suggestion that, whenever possible, analyses at the individual level should be done as well as group analyses (Estes & Maddox, 2005). At the very least, the level of heterogeneity within the considered sample should be explicitly evaluated (Smith & Batchelder, 2008, 2010). Nevertheless, this approach is not always possible (or advisable) because there are sometimes too few data from each individual, or only a very small number of individuals are considered, or they are chosen from a special population. Under those circumstances, group fitting could be the only possible strategy for analysis (Cohen, Sanborn, & Shiffrin, 2008). In the same vein, models of medium-high complexity should generally be fitted at the individual level (Estes & Maddox, 2005).

When experimental levels are manipulated between-subjects, we cannot decide in general whether the effects are present in a specific individual. In those instances, group fitting must be employed. But when individual fitting is possible, it allows assessing whether individuals show the same effects as the group as a whole. Even when a single model exists and it is properly defined, there are advantages over group-fitting that can be derived from fitting data at the individual level (Estes & Maddox, 2005).

General and differential psychological approaches work better together (Grandy et al., 2017; Lamiell, 2003; Molenaar & Campbell, 2009; Nesselroade, 2010; Voelkle et al., 2014). Understanding human behavior is the key goal shared by both disciplines, and we should apply all our best weapons to achieve this goal. This is not about which approach is more favorable in any circumstance, but about how to combine them judiciously to obtain reliable answers to the questions of interest.

5. REFERENCES

Arend, I., Colom, R., Botella, J., Contreras, M. J., Rubio, V. y Santacreu, J. (2003). Quantifying cognitive complexity: evidence from a reasoning task. *Personality and Individual Differences*, 35, 659-669.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624-652.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulated cortex: An update. *Cognitive Sciences*, *8*(12), 539- 546.

Broadbent, D. E. (1958). *Perception and Communication.* London, Pergamon.

Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97,* 404-431.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*(4), 692-712.

Colom, R. (2016). Advances in intelligence research: What should be expected in the XXI Century? (Questions and Answers). *Spanish Journal of Psychology*, 19. E92. https://doi.org/10.1017/sjp.2016.87

Colom, R., Abad, F. J., Quiroga, Mª A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence, 36,* 584-606.

Colom, R., Chuderski, A., & Santarnecchi, E. (2016). Bridge over troubled water: Commenting on Kovacs & Conway's Process Overlap Theory. *Psychological Inquiry, 27, 3,* 181-189.

Cooper, L. (1982). Strategies for visual comparison and representation: Individual differences. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence,* Vol. 1. LEA.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116-127.

Danziger, K. (1990). *Constructing the Subject: Historical Origins of Psychological Research*. Cambridge: Cambridge University Press.

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist, 70,* 499-514.

Deutsch, J. A., & Deutsch, D. (1963). Attention: some theoretical considerations. *Psychological Review, 70:* 80-90.

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20, 6,* 425-443.

Duthoo, W., Abrahamse, E. L., Braem, S., Boehler, C. N., & Notebaert, W. (2014). The heterogenous world of congruency sequence effects: An update. *Frontiers in Psychology,* 5, 1001. doi: 10.3389/fpysg2014.01001.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53(2),* 134-140. doi: 10.1037/h0045156

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12*(3), 403-408.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M., Papademtris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience, 1*8, 1664-1671.

Gigerenzer, G. (1987). Survival of the fittest probabilist: Brunswik, Thurstone, and the two disciplines of psychology. In L. Krüger, G. Gigerenzer & M. S. Morgan (Eds.), *The Probabilistic Revolution: Ideas in the Sciences* (Vol. 2, pp. 49-72). Cambridge, MA: MIT Press.

Grabot, L., & van Wassenhove, V. (2017). Time order as psychological bias. *Psychological Science,* 28, 670-678. doi: 10.1177/0956797616689369.

Grandy, T. H., Lindenberger, U., & Werkle-Bergner, M. (2017). When group means fail: Can one size fit all? bioRxiv, 126490. https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1101%2F126490&data=02%7C01%7Cjuolas%40ku.edu%7C15172d325a2c444b7b3a08d655c6a31b%7C3c176536afe643f5b96636feabbe3c1a%7C0%7C1%7C636790708967332217&sdata=GX9jxRvBAleMxXShjLJOY68A6whCsbWN6tgbNrReH8w%3D&reserved=0

Haier, R. J. (2017). *The Neuroscience of Intelligence*. Cambridge, UK: Cambridge University Press.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*, 494-500.

Hommel, B. (2007). Feature integration across perception and action: Event files affect response choice. *Psychological Research, 71*, 42-63.

Hommel, B., Proctor, R. W., & Vu, K. P. L. (2004). A feature-integration account of sequential effects in the Simon task. *Psychological Research, 68*, 1-17.

Hunt, E. B. (1978). Mechanics of verbal ability. *Psychological Review, 85*, 109-130.

Lamiell, J. T. (2003). *Beyond Individual and Group Differences: Human Individuality, Scientific Psychology, and William Stern's Critical Personalism*. Thousand Oaks, CA: Sage.

Lavie, N. (2001). Capacity limits in selective attention: Behavioral evidence and implications for neural activity. In J. Braun, C. Koch, & J. Davis (Eds.). *Visual Attention and Cortical Circuits*. Cambridge, MA: MIT Press.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605-621.

Luck, S. J., & Vogel, E. K., 2013. Visual working memory capacity: From psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences, 17*, 8,391-400.

Martínez, K., Burgaleta, M., Roman, F. J., Escorial, S., Shih, P. C., Quiroga, M. A., & Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence, 39*, 473-480.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* Belmont, CA: Wadsworth.

Mayr, U., Awh, E., & Laurey, P. (2003). Conflict adaptation effects in the absence of executive control. *Nature Neuroscience, 6*(5), 450-452).

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, 18(2), 112-117. doi: 10.1111/j.1467-8721.2009.01619.x

Nesselroade, J. R. (2010). On an emerging third discipline of scientific psychology. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development* (pp. 209-218). Washington, DC: American Psychological Association; US.

Open Science Collaboration (2015). *Science*, 349, aac4716. doi: 101126/science.aac4716.

Privado, J., Botella, J., Suero, M., Quiroga, M. A., & Colom, R. (2015). Still seeking for an explanation of the Sequential Compatibility Effect. *Anales de Psicología*, 31(2), 687-696.

Rapp, B. (2012). Case series in cognitive neuropsychology: Promise, perils, and proper perspective. *Cognitive Neuropsychology*, 28(7), 435-444. doi: 10.1080/02643294.2012.697453.

Simon, J. R. (1969). Reactions towards the source of stimulation. *Journal of Experimental Psychology, 81*, 174-176.

Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. In R.W. Proctor & T.G. Reeve (Eds.). *Stimulus-response Compatibility*. Amsterdam: Elsevier.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review, 15*(4), 713-731.

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*(1), 167-183.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, 1-19.

Sternberg, R. J. (1977). *Intelligence, information processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R. J. (1979). Intelligence research and the interface between differential and cognitive psychology: prospects and proposals. In R. J. Sternberg and D. K. Detterman (Eds.), *Human intelligence: Perspectives on its theory and measurement.* New York: Ablex.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology: General, 109*, 119-159.

Sternberg, S. (1966). High-speed scanning in human memory. *Science,* 153, 652-654.Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention». *American Journal of Psychology, 77,* 206-219.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71*, 1-26.

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research,* 49(3), 193-213. doi: 10.1080/00273171.2014.889593

TABLE 1. Results of the ANOVAs (main effects of congruency of the current trial and the interactions with congruency of the previous trial) for the three samples.

Sample	Congruency current trial	Interaction
A	$F(1,377) = 553.2$; $p < .001$	$F(1, 377) = 711.6$; $p < .001$
B	$F(1,477) = 1{,}008.0$; $p < .001$	$F(1, 477) = 782.5$; $p < .001$
C	$F(1,416) = 708.0$; $p < .001$	$F(1, 416) = 412.5$; $p < .001$

TABLE 2. Results of the ANOVAs (main effects of the congruency of the current trial and the interactions with the congruency of the previous trial) for the three samples after excluding the trials that are pure replicas.

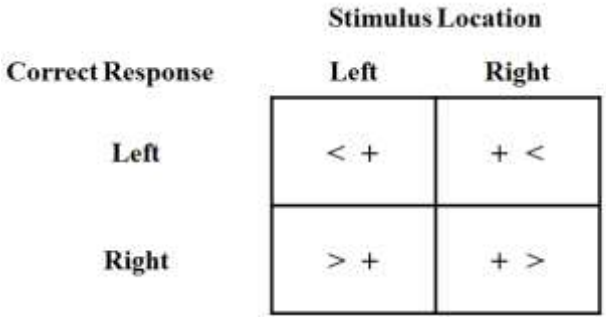
Sample	Congruency current trial	Interaction
A	$F(1,377) = 441.3$; $p < .001$	$F(1,377) = 259.8$; $p < .001$
B	$F(1,476) = 739.5$; $p < .001$	$F(1,477) = 142.0$; $p < .001$
C	$F(1,416) = 479.4$; $p < .001$	$F(1,416) = 113.2$; $p < .001$

TABLE 3. Results classifying individuals according to the rules specified in the text for the three samples (N_A = 379, N_B = 358, N_C = 422)

Models	**Sample A**	**Sample B**	**Sample C**
Model 0	184	150	140
Model 1	82	92	146
Model 2	22	28	49
Models 2 & 3	34	26	31
Model 3	23	40	37
Model 4	34	22	19

FIGURE 1. Experimental paradigm analyzed in the present study. The fixation cross always appeared in the center of the display. In Figure 1(a) congruent trials are those in the upper left and lower right quadrants. In Figure 1(b) both congruent and incongruent trials can be preceded by identical, replica trials or by congruent or incongruent trials with different stimulus arrangements.

(a)



(b)

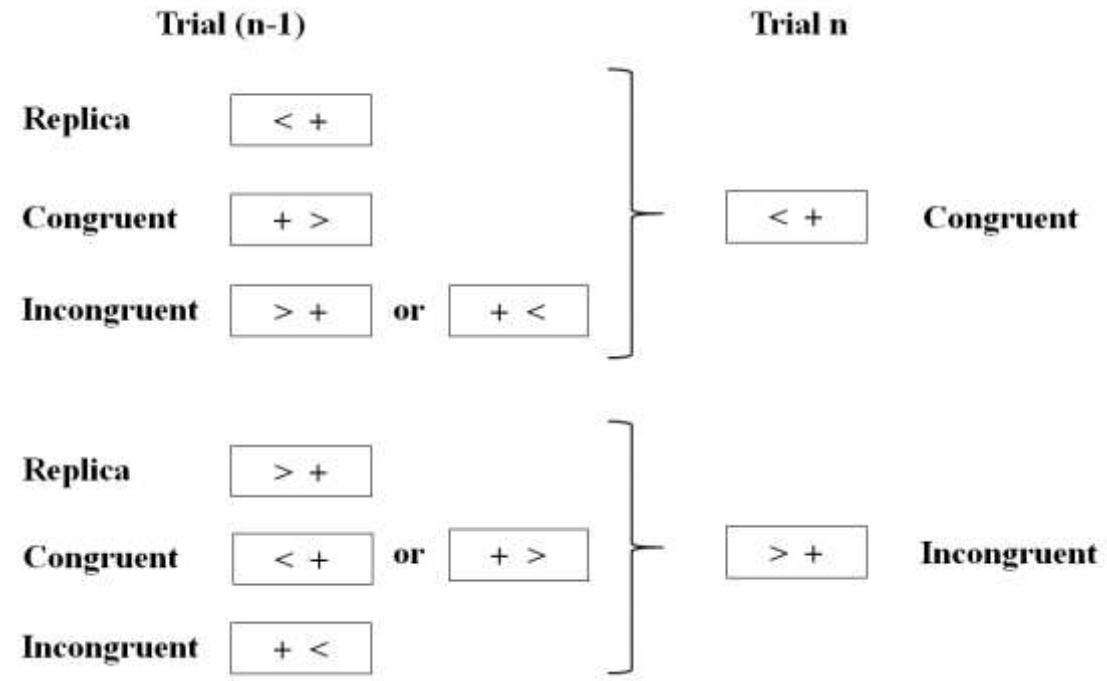


FIGURE 2. Schematic representation of the rules for classifying each individual in one of the models assessed

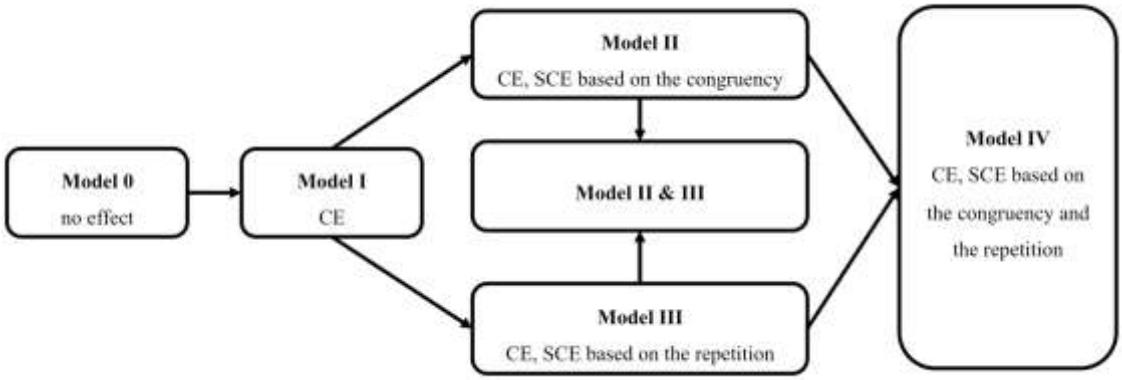


FIGURE 3. Mean RTs for the two conditions of congruency on trial *N* as a function of the congruency of the previous trial (*N*-1); (a) with all the trials (left) and (b) after eliminating the pure replicas (right).

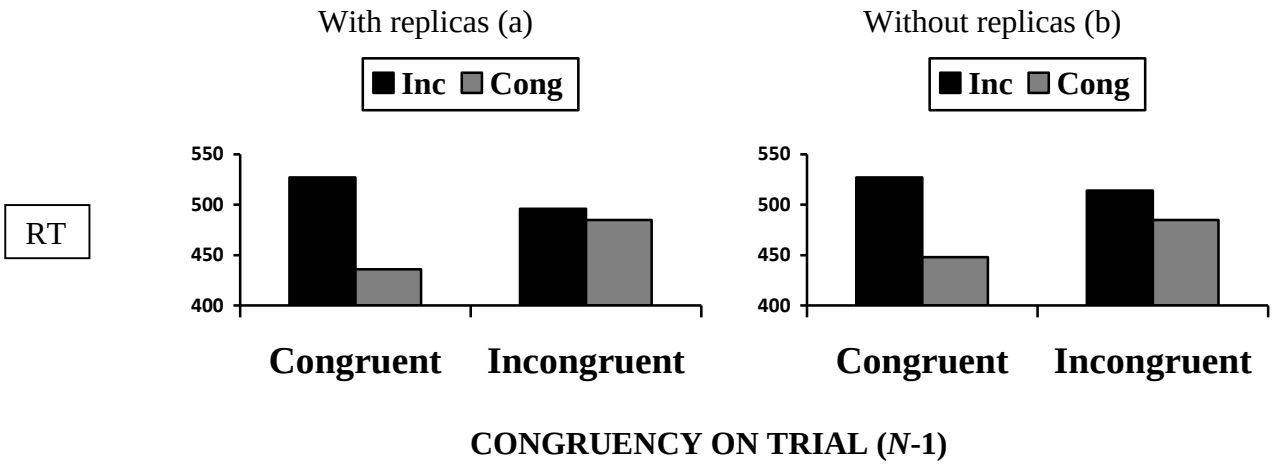


FIGURE 4. Summary showing the percent of participants classified within the models depicted in Figure 2 and described in the main text (N = 1,159).