PSYCHOMETRIC MODELS OF SMALL GROUP COLLABORATIONS

Peter F. Halpin and Yoav Bergner
new york university

Correrspondence:
Peter Halpin
246 Greene Street, Office 204
New York, New York
10012
212 998 5187
peter.halpin@nyu.edu

Abstract

The social combination theory of group problem solving is used to extend existing psychometric models to collaborative settings. A model for pairwise group work is proposed, the implications of the model for assessment design are considered, and its estimation is addressed. The results are illustrated with an empirical example in which dyads work together on a twelfth-grade level mathematics assessment. In concluding, attention is given to specific avenues of research that seem most fruitful for advancing current initiatives concerning the assessment of collaboration, team work, and related contructs.

Key words: group work, collaborative problem solving, item response theory, social combination theory, process loss

1. Introduction

Despite a relatively long history of research on small group collaborations in assessment settings (Webb, 1995), there has not yet been an attempt to formulate a psychometric theory of collaboration. The need for such a theory is underscored by recent initiatives concerning the assessment of collaborative problem solving, team work, and related constructs (e.g., von Davier, Kyllonen, & Zhu, 2017; Heckman & Kautz, 2014; Herman & Hilton, 2017; Fiore et al., 2017; Griffin & Care, 2015; Lippman, Ryberg, Carney, & Moore, 2015; National Research Council, 2011; Organisation for Economic Co-operation and Development, 2013; Pellegrino & Hilton, 2012; Stecher & Hamilton, 2014). In this paper, it is argued that the social combination approach to group problem solving (see e.g., Laughlin, 2011, chap. 2) provides a suitable framework for extending existing psychometric models to collaborative settings.

The literature on social combination theory is selectively reviewed to motivate an overall modeling framework. We then propose a model of pairwise group work, which leads to a number of results about the design of group assessments. In particular, we consider how to select items and group members such that the expected performance of a dyad can be empirically distinguished from that of either individual. We then address estimation of the model, and the results are illustrated with data simulation and an empirical example in which pairs of respondents work together to complete a twelfth-grade level mathematics assessment.

The paper has a number of limitations which can be mentioned at the outset. We focus on binary (correct/incorrect) response data, unidimensional latent traits for individual performance, and pairwise group work in which the groups are assumed to be independent from one other.

Additionally, due to the nature of the example data, this research has so far been limited to inferential methods that can be applied when only a relatively small number of dyads are observed working together (e.g., we do not estimate item parameters under group testing conditions). In the concluding section, these limitations are discussed in terms of specific avenues of future research that seem most fruitful for advancing current initiatives concerning the assessment of collaboration, team work, and related domains.

2. Social Combination Models of Group Problem Solving

Lorge and Solomon (1955) were among the first to use statistical modeling to study group problem solving, and the following quotation provides a germane starting point for the present research.

> Under Model A the probability of a group solution is the probability that the group contains one or more members who can solve the problem. This non-interactional ability model for any specific problem can be expressed mathematically as follows: Let
>
> P_G = the probability that a group of size k solve the problem;
>
> P_I = the probability that an individual solve the problem.
>
> Then
>
> $$P_G = 1 - (1 - P_I)^k$$
>
> where P_G and P_I are population parameters considered fixed for the specific problem and the specific population (Lorge & Solomon, 1955, p. 141).

Under Model A, the probability that a group solves a problem is just the probability that at least one of its members solves the problem. Lorge and Solomon discuss the conditions under which their model is plausible, but for now we may simply observe that Model A provides a clear point of contact between psychometrics and small group research. By replacing P_I with an appropriate item response theory (IRT) model, P_G can be treated as conditional on features of the problem and of the respondents. Social combination models such as Lorge and Solomon's Model A can then be re-interpreted as item response functions for groups of respondents working together on an educational or psychological assessment. This is the basic idea behind the present research.

A large number of statistical models of problem solving and decision making have been proposed since Lorge and Solomon's (1955) paper. The rest of this section outlines the main features of the approach taken in the present research. Later on, we will revisit a psychometric re-formulation of Model A.

2.1. Task Types

It is useful to begin by delineating the types of tasks that are under consideration. McGrath's (1984) circumplex typology distinguishes tasks according to two theoretical dimensions. One dimension represents a contrast between tasks that incentivize either cooperation or conflict among group members. The second dimension represents a contrast between tasks that are principally cognitive versus behavioral in nature. *Intellective tasks* (Laughlin, 1980) are located in the cooperative-cognitive quadrant of the circumplex, and are characterized by problem-solving scenarios in which there exists a demonstrably correct answer. This type of task is exemplified by problems in mathematics and logic, as well as problems that are about factual content. Lorge and Solomon (1955) were concerned with a subset of intellective tasks, so-called “Eureka-type” problems, in which the correct answer is assumed to be immediately recognized as such when it is revealed to an individual.[1] Clearly, assessments that are scored using a correct/incorrect scheme fall under the rubric of intellective tasks. Therefore, intellective tasks provide the focal point of the present research.

In contrast to intellective tasks, *decision-making* tasks are problem-solving scenarios in which no agreed-upon correct answer exists (e.g., Davis, 1992). This type of task is exemplified by jury deliberation, and it typically involves some degree of conflict among group members. The distinction from intellective tasks can be quite fuzzy, especially for intellective tasks that are not of the Eureka type (e.g., in cases where team members are prone to disagree about the correctness of a solution). Although related, this type of task is not the central focus of the present research. It is also useful to contrast intellective tasks with *mixed-motive* tasks, which are exemplified by the prisoner’s dilemma (e.g., Poundstone, 1992). Mixed-motive tasks incentivize individuals to strategize and act against one another’s best interests. On the other hand, intellective tasks incentivize group members to work towards the same, shared goal – i.e., to provide a correct response. In McGrath’s typology, these two types of tasks occupy opposite quadrants of the circumplex.

2.2. Defining Correct Responses in a Group Setting

We have narrowed the focus to tasks that require individuals to cooperate in order to provide demonstrably correct responses. This leads to the question of what constitutes a correct response

[1] Examples include the Tower of Hanoi and the Tartaglia – Laughlin (2011, chap. 2) provides details.

in a group setting. Here it is useful to consider the distinction between *unitary* and *divisible* tasks (Steiner, 1972). A unitary task is one that requires a single result or product from all group members, whereas divisible tasks permit different output from different subsets of group members. In the context of assessment, this distinction can be interpreted in terms of how a task is scored. For example, we may wish to score the group as a whole, or the responses of individual members separately. While these objectives need not be incompatible, the focus of the present research is to evaluate the group as a whole.

The notion of a group response can be formalized in terms of scoring rules that are applied to the responses of individual members. The types of unitary tasks identified by Steiner (1972) provide a number of plausible scoring rules. For example, when using a disjunctive scoring rule, a group is regarded as producing a correct result just in case any of its members do. Conversely, a conjunctive scoring rule defines a group's response as correct just in case all of its members provide a correct response. A number of psychometric models make similar distinctions, such as compensatory versus non-compensatory multidimensional item response theory models (e.g., Reckase, 2009), and "noisy-and" versus "noisy-or" models in cognitive diagnostic assessment (e.g., Junker & Sijtsma, 2001). As an intuitive starting point, this research focusses on conjunctive scoring rules.

2.3. Modeling Group Performance

A key ingredient in social combination theory is the *decision function*, which was introduced in the following quotation from Smoke and Zajonc (1962, p. 322):

> If p is the probability that a given individual member is correct, the group has a

> probability $h(p)$ of being correct, where $h(p)$ is a function of p depending upon the type of decision scheme accepted by the group. We shall call $h(p)$ a decision function.

The decision function is a generalization of Lorge and Solomon's (1955) Model A that characterizes the types of models used in social combination theory. Research in this area has typically focussed testing theoretically motivated decision functions against data aggregated over groups (see, e.g., Laughlin, 2011). Davis (1973) was among the first to address estimation of parametric decision functions from group response data, and his general model is a main feature of the approach developed below.

It is important to keep in mind the distinction between scoring rules and decision functions – whereas the former is a feature of the assessment decided by the test designer, the latter characterizes the behavior of groups during the test. The two are related in the sense that groups may adopt different decision schemes in order to optimize their performance with respect to different scoring rules. However, this game-theoretic aspect of test design is not addressed in the present paper.

2.4. Summary

The basic goal of the current research is to develop a psychometric theory for tasks in which group members cooperate to provide demonstrably correct responses. A group's responses are defined in terms of scoring rules that can be applied to the responses of its individual members, and the expected performance of a group is modeled in terms of the expected performance of the individual group members.

3. A psychometric Approach to Group Performance

To provide a psychometric framing of social combination theory, we begin by specifying assessments administered to individual respondents. Next we define group assessments such that they include individual assessments as the special case where all groups have a single member. Finally, we use Davis' (1973) general social combination (GSC) model to extend existing psychometric models for individual assessments to cases where groups have multiple members.

3.1. Individual Assessments

For item $i = 1, \dots, I$ of a test T, and respondent $j = 1, \dots, J$, let

$$X_{ij} = \begin{cases} 1 & \text{if respondent } j \text{ answers item } i \text{ correctly.} \\ 0 & \text{if respondent } j \text{ answers item } i \text{ incorrectly.} \end{cases} \tag{1}$$

In this paper we stipulate that the responses $\boldsymbol{X}_j = (X_{1j}, \dots, X_{Ij})$ can be described in terms of a univariate monotone latent variable (UMLV) model (Holland & Rosenbaum, 1986, sections 2.1-2.3). Assuming the existence of a latent variable $\theta \in \mathbb{R}$, a UMLV model is defined via the following two conditions on the joint distribution of the response vector $\boldsymbol{X}_j$ and θ_j. First, the item responses $\boldsymbol{X}_j$ are conditionally independent given θ_j:

$$F(\boldsymbol{X}_j \mid \theta_j) = \prod_{i=1}^{I} F_i(X_{ij} \mid \theta_j). \tag{2}$$

The second requirement is latent monotonicity of the item response functions (IRFs), which, for binary responses, can be stated in terms of the expected value of X_{ij} given θ_j:

$$E(X_{ij} \mid \theta_j) \leq E(X_{ij} \mid \theta_j'), \tag{3}$$

for $\theta_j < \theta'_j$ and $i = 1, \dots, I$. Below we make use the notation $P_{ij} = P_i(\theta_j) = E(X_{ij} \mid \theta_j)$ and $Q_{ij} = 1 - P_{ij}$.

3.2. Group Assessments

We consider the case where groups are formed by assignment without replacement of respondents to groups. Let the set $\mathcal{J} = \{j[1], \dots, j[J]\}$ denote the respondents and G denote a partition of $\mathcal{J}$ with elements $G_k = \{j[k_1], \dots, j[k_n]\}$ that satisfy

$$G_k \cap G_{k'} = \emptyset, \quad k \neq k', \quad \text{and} \quad \cup_{k=1}^{J/n} G_k = \mathcal{J}.$$

Then n is the number of respondents in group k, here assumed to be constant, and $K = J/n$ is the number of non-overlapping groups formed from a pool of J respondents.

For item $i = 1, \dots, I$ of a test T', let

$$Y_{ik} = \begin{cases} 1 & \text{if group } k \text{ answers item } i \text{ correctly.} \\ 0 & \text{if group } k \text{ answer item } i \text{ incorrectly.} \end{cases} \tag{4}$$

As mentioned, the focus of the present paper is group responses that result from applying a conjunctive scoring rule to the responses of individual group members, say $Y_{ik} = \prod_{r \in G_k} Y^*_{ir}$, with Y^*_{ir} denoting the individual responses. However, the specification of a group response in Equation (4) is intended to capture any type of correct / incorrect group scoring procedure.

Similar to the UMLV model for individual assessments, we will be interested in models where the responses $\boldsymbol{Y}_k = (Y_{1k}, \dots, Y_{Ik})$ are conditionally independent given $\boldsymbol{\theta}_k = (\theta_{k_1}, \dots, \theta_{k_n})$,

$$F(\boldsymbol{Y}_k \mid \boldsymbol{\theta}_k) = \prod_{i=1}^{I} F_i(Y_{ik} \mid \boldsymbol{\theta}_k). \tag{5}$$

Given this assumption, it will be sufficient to specify a model for $\boldsymbol{Y}_k$ in terms of the group IRFs, denoted as $R_{ik} = R_i(\boldsymbol{\theta}_k) = E(Y_{ik} \mid \boldsymbol{\theta}_k)$.

When $n = 1$, we stipulate that $R_{ik} = P_{ik}$ – i.e., that the group IRF is just an individual IRF, as defined in Equation (3). Consequently, when $n = 1$, the definition of group assessments in Equations (4) and (5) reduces to that of individual assessments given in Equation (1) through (3). In more practical terms, we are assuming that each group assessment has a corresponding individual assessment that is identical other than the instructions pertaining to how respondents may work together. We refer to this as the individual version of a group assessment. When $n > 1$, the goal is to derive the properties of R_{ik} from the P_{ir}, $r \in G_k$, using the following model for group performance.

3.3. A General Social Combination Model

Davis (1973) presented a general formulation of social combination models that we adapt to the present context as follows. Let $\boldsymbol{X}^*_{ik} = (X_{ik_1}, \ldots, X_{ik_n})$ denote the responses that would have resulted if each member of group k had written the individual version of item i on a group assessment T'. There are $S = 2^n$ possible realizations of $\boldsymbol{X}^*_{ik}$, each with probability

$$\pi_{iks} = \text{Prob}(\boldsymbol{X}^*_{ik} = x_s \mid \boldsymbol{\theta}_n) = \prod_{r=1}^{n} P_{ik_r}^{\,x_{ik_r}} \, Q_{ik_r}^{\,1-x_{ik_r}}, \quad s = 1, \ldots, S$$

following directly from the properties of individual assessments. Note that π_{iks} is the probability of n responses to a single item i, not to be mistaken with the more familiar expression for I responses of a single individual.

For the binary group response defined in Equation (4), a social combination model can be specified as $2 \times S$ matrix $D_k = \{d_{rs}\}$ that maps the probabilities of the individual responses, $\boldsymbol{\pi}_{ik} = (\pi_{ik1}, \ldots, \pi_{ikS})$, onto the probabilities of the group responses $\boldsymbol{\rho}_{ik} = (R_{ik}, 1 - R_{ik})$. The general model to be considered is then

$$\boldsymbol{\rho}_{ik} = D_k \, \boldsymbol{\pi}_{ik} \, , \tag{6}$$

in which it is required that $d_{rs} \in [0, 1]$ and $\sum_r d_{rs} = 1$, for each $s = 1, \ldots, S$, to ensure that $R_{ik} \in [0, 1]$.

In comparison with the model considered by Davis (1973), Equation (6) has that following advantages: (a) it does not assume that the probability of a correct response to an item is equal for all individuals, (b) it allows these probabilities to vary over items, and (c) it allows the decision function D_k to vary over groups. It may therefore be interpreted as a psychometric re-formulation of social combination theory.

4. Models for Pairwise Group Performance

In the remainder of this paper we confine attention to pairwise group work. To simplify notation we write $k_r = r$ for $r = 1, 2$ and drop the subscript k for groups. We will focus on the following restricted social combination (RSC) model:

$$\begin{bmatrix} R_i \\ 1 - R_i \end{bmatrix} = \begin{bmatrix} 1 & a & b & 0 \\ 0 & 1-a & 1-b & 1 \end{bmatrix} \times \begin{bmatrix} P_{i1}P_{i2} \\ P_{i1}Q_{i2} \\ Q_{i1}P_{i2} \\ Q_{i1}Q_{i2} \end{bmatrix} , \tag{7}$$

in which the zeros of the decision function matrix are considered to be structural constraints. We propose the RSC model as a starting point for studying group assessments that use a conjunctive scoring rule, while recognizing that many applications may require more general models.

The focus of this section is the interpretation of four special cases of the RSC model that arise by setting $a, b \in \{0, 1\}$, with the resulting group IRFs denoted as R_i^{ab}. In the next section the relations among these four models are discussed from the perspective of assessment design. Subsequently, we turn to address the estimation non-integer values of a and b from group response data. Before moving on, we note the following property of the RSC model, which provides a starting point for its interpretation.

Proposition 1. Let $R_i(\boldsymbol{\theta})$ denote a group IRF obtained from the RSC model in Equation (7) with $a, b \in [0, 1]$. If $P_i(\theta)$ satisfies latent monotonicity (i.e., is monotone non-decreasing in θ; see Equation (3)), then $R_i(\boldsymbol{\theta})$ also satisfies latent monotonicity, in each coordinate of $\boldsymbol{\theta}$.

The proof of Proposition 1 is in the Appendix.

4.1. The Independence Model

Setting $a = b = 0$, we obtain the group IRF $R_i^{00} = P_{i1}P_{i2}$. Under the conjunctive scoring rule for Y_i, this describes the case in which the expected performance of the group is equivalent to what would be expected from the respondents had they worked independently. We therefore refer to R_i^{00} as the Independence model. The structural zero in the first column of the decision function in Equation (7) implies that the IRF of the Independence model is a lower bound on IRFs obtained using any values of $a, b \in [0, 1]$. Otherwise stated, the RSC model implies that groups do not perform worse than their individual members working independently. Given appropriate instructions for group assessments, we expect that the Independence model would be a reasonable lower bound on empirical group performance. For example, respondents might be instructed that they may choose to work without their partners at any point during the test.

4.2. Individual Performance

Next consider two models in which the expected performance of the group reduces to that of either individual. Without loss of generality we assume that $\theta_1 \leq \theta_2$ and write

$$R_i^{10} = P_{i1} \quad \text{and} \quad R_i^{01} = P_{i2}, \tag{8}$$

which we will refer to as the "Min model" and the "Max model", respectively. As described by Webb (1995), various kinds of participation biases can lead to this kind of group performance. Regarding the Max model, it can be both efficient and effective for groups to defer output to the most capable individual, when a task is intended to measure group productivity. The Min model may arise when group members' participation is influenced by status characteristics that are not related to their ability. Such participation biases are more likely to occur when group members do not have a clear way to judge each other's competence on the task, or in pre-existing groups where the relative status of group members has been established in contexts outside of the task (Webb, 1995). In short, if one member of a group is exclusively responsible for providing responses, then the group would be expected to perform at the level of that individual.

As mentioned in the introduction, we are especially interested in designing group assessments such that the expected performance of a dyad can be empirically distinguished from that of either individual. Thus, in addition to their empirical interpretations in terms of group participation bias, these models for individual performance will serve as important reference points for designing group assessments.

4.3. Better-than-Individual Performance

In many contexts, the following intuition about group work arises:

$$E(Y_i \mid \boldsymbol{\theta}) \geq R_i^{01},$$

which we refer to as better-than-individual performance. As an obvious example of better-than-individual performance, consider a situation in which group members have differential access to the information required to solve a problem. Social psychologists often induce this type of situation experimentally (e.g., Stasser & Titus, 2003), it is a basic premise of team work in organizational settings (e.g., Mesmer-Magnus & DeChurch, 2009; Salas, Cooke, & Rosen, 2008), and its importance has been consistently underscored by educational theories of group work (e.g., Cohen, Lotan, Abram, Scarloss, & Schultz, 2002; Aronson, Blaney, Stephan, Sikes, & Snapp, 1978). In the context of educational testing, it is not difficult to imagine that two respondents with complementary skill sets would be expected to produce more correct responses than either person working in isolation.

One way to model better-than-individual performance is in terms of an Additive model:

$$R_i^{11} = P_{i1}\,Q_{i2} + Q_{i1}\,P_{i2} + P_{i1}\,P_{i2} = 1 - Q_{i1}\,Q_{i2}. \tag{9}$$

This Additive model is a re-formulation of Lorge and Solomon's (1955) Model A. The structural zero in the last column of the decision function in Equation (7) implies that R_i^{11} is an upper bound on the IRFs of the RSC model. This interpretation of the Additive model as an upper bound on group performance is supported by a large number of experimental studies showing that group performance on intellective tasks very rarely exceeds the level predicted by Lorge and Solomon's Model A (see reviews by Steiner, 1972; McGrath, 1984; Davis, 1992). Steiner (1972) also provided a theoretical rationale, based on information sharing, that supported the interpretation of Model A as the ideal or maximal group performance on intellective tasks (his "truth-wins" criterion). On this interpretation, values of $a, b < 1$ correspond to what Steiner termed *process loss*, which describes the discrepancy between a group's theoretical maximum performance and its actual performance. The standard examples of process loss in the social psychology literature include lack of motivation of one or more members and inefficient coordination of activities among group members. More recently, literature on learning and assessment has described many individual and group attributes that are theorized to lead to successful collaborations (e.g., Fiore et al., 2017; Griffin & Care, 2015; Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015), and these may also be interpreted as potential sources of process loss for appropriately designed tasks.

It is important to emphasize that the interpretation of the Additive model as an upper bound on group performance is based on empirical and theoretical considerations, but lacks a mathematical derivation. We address the applicability of the RSC model to group assessment data in our empirical example.

5. Model Equivalence and Assessment Design

The four models considered above are related as follows

$$R_i^{00} \leq R_i^{10} \leq R_i^{01} \leq R_i^{11}. \tag{10}$$

The reader may have already noted a number of conditions under which these inequalities can be replaced by strict equalities. Most obviously, if $\theta_1 = \theta_2$ then $R_i^{10} = R_i^{01}$ for all i. Using the terminology of Vuong (1989, Def. 3), the Min model and the Max model are *overlapping*, because they imply the same distribution of $\boldsymbol{Y}$ for some values of $\boldsymbol{\theta}$, but neither model is nested within the other. Clearly, in cases where two models imply the same distribution for $\boldsymbol{Y}$, it will not be possible to distinguish the models with empirical data. We therefore refer to situations in which the proposed models overlap as a problem of model equivalence.

This section considers how to design group assessments so as to avoid equivalence among the four models. Concerning the Min and Max models, it is apparent that we should chose group members such that $\theta_1 \neq \theta_2$. Given this requirement on team composition, it will then possible to select items such that $R_i^{10} < R_i^{01}$, and model equivalence is thereby avoided. In particular, the difference $R_i^{01} - R_i^{10}$ will be large for highly discriminating items that are targeted at the level of the more able respondent.

In addressing the other inequalities in Expression (10), our overall concern is to ensure that a group assessment is designed such that it can distinguish the expected performance of a group from that of its individual members. As we now show, this is a non-trivial consideration. All proofs are contained in the Appendix unless otherwise noted. As above, we assume that $\theta_1 \leq \theta_2$ by choice of notation.

5.1. Some Results on the Design of Group Assessments

The following proposition shows that the equivalence of the Additive model and the Max model, and of the Min model and the Independence model, are in fact the same problem.

Proposition 2. For group members $1, 2$ and item i, let $\Delta_{i12} = P_{i1}\,Q_{i2}$. Then

$$\Delta_{i12} = R_i^{10} - R_i^{00} = R_i^{11} - R_i^{01}. \tag{11}$$

Proposition 2 follows immediately from the definitions of the IRFs for the four models. It implies that analysis of Δ_{i12} for an arbitrary item and group is sufficient to describe equivalence between Additive and Max models, as well as between the Min and Independence models. The quantity Δ_{i12} will be referred to as the *item delta* for item i and group members 1 and 2.

Clearly, $\Delta_{i12} = 0$ just in case $P_{i1} = 0$ or $Q_{i2} = 0$. The first case occurs when the less able partner is certain to provide an incorrect response, when working individually. The second case occurs when the more able partner is certain to provide a correct response, when working individually. These two situations correspond to what Shiflett (1979) described as *redudancy* of team resources. When considering Δ_{i12} as a function of i, we will refer to task or item redundancy. When considering it as a function of θ_r, $r = 1, 2$, we will refer to redundancy of team members. To avoid confusion with established uses of the term "information" in IRT, we do not use it as an antonym for redundancy. For further interpretations of redundancy in team settings, see Mesmer-Magnus and DeChurch (2009).

Proposition 3 addresses how to avoid redundancy when designing group assessments.

Proposition 3. Let $\Delta(P_{i1}, P_{i2})$ denote the item delta in Proposition 2, treated as a function of P_{i1} and P_{i2}. The requirement $0 < P_{i1} \leq P_{i2} < 1$ implies that $\Delta(P_{i1}, P_{i2})$ is strictly concave, with global maximum $\Delta(1/2, 1/2) = 1/4$.

The following result describes a special case of Proposition 3 that applies to many IRT models of binary data.

Proposition 4. Let $\Delta_i(\boldsymbol{\theta})$ denote the item delta in Proposition 2, treated as a function of $\boldsymbol{\theta} = (\theta_1, \theta_2)$. If $P_i(\theta)$ is strictly increasing in a neighbourhood $\mathcal{N}$ around $\theta_i^* = \{\theta \mid P(\theta) = .5\}$, then

Part 1. $\arg\max_{\boldsymbol{\theta}} \{\Delta_i(\boldsymbol{\theta})\} = (\theta_i^*, \theta_i^*)$.

Part 2. For $\theta_1 \leq \theta_i^* \leq \theta_2 \in \mathcal{N}, \;\; \Delta_i(\boldsymbol{\theta})$ is strictly decreasing with $\delta = \theta_2 - \theta_1$.

Part 1 of Proposition 4 states that item redundancy will be minimized when both group members have ability equal to the difficulty of the item. Moreover, Part 2 states that item redundancy is strictly increasing as the ability level of either partner moves away from the difficulty level of the item. Consequently, it is apparent that a group assessment designed to avoid redundancy will have different characteristics than one designed to distinguish between the Min and Max models. We revisit this point in the section summary.

In practice, it will not generally be feasible to select partners and items to satisfy Part 1 of Proposition 4, especially when the pool of examinees and / or items is quite small. To address this situation, the remainder of this section describes the problem of item selection for cases where $\delta \geq 0$. However, all of the following results also require stronger assumptions on the UMLV model for individual performance, which are stated as part of the following proposition.

Proposition 5. Assume that the IRFs for individual performance can be written as a two-parameter logistic (2PL) model

$$P(\theta) = [1 + \exp\{-\alpha_i(\theta - \beta_i)\}]^{-1} \tag{12}$$

with $\beta_i \in \mathbb{R}$ and $\alpha_i > 0$. Let $\Delta_{12}(\beta_i)$ denote the item delta in Proposition 2, treated as a function of β_i for any fixed values of $\theta_1 \leq \theta_2$. Then

$$\beta_i^* \equiv \arg\max_{\beta_i} \{\Delta_{12}(\beta_i)\} = (\theta_1 + \theta_2)/2.$$

and $\Delta_{12}(u)$ is monotone decreasing in $u = |\beta_i - \beta_i^*|$.

Although the assumption of a 2PL IRF is quite restrictive, the result is nonetheless interesting because of its simplicity. For any pair of respondents, items chosen to have difficulty equal to the average of the group members' abilities will be least redundant, and items will become increasingly redundant the farther they move away from this optimal value.

The next proposition addresses the role of the discrimination parameter.

Proposition 6. Assume that the IRFs for individual performance can be written as in Proposition 5 and let $\Delta^*_{12}(\alpha_i)$ denote the item delta in Proposition 2 treated as a function of α_i, evaluated at β_i^*, for any fixed values of $\theta_1 \leq \theta_2$. Then for $\alpha_i < \alpha'_i$,

$$\Delta^*_{12}(\alpha_i) \geq \Delta^*_{12}(\alpha'_i).$$

When $\theta_1 = \theta_2$ the inequality can be replaced with an equality. Otherwise, $\Delta^*_{12}(\alpha_i) \rightarrow 0$ as $\alpha_i \rightarrow \infty$.

Proposition 6 shows that the minimal item redundancy for fixed values of $\theta_1 < \theta_2$ is decreasing in the item discrimination. On the other hand, when $\theta_1 = \theta_2$, $\Delta^*_{12}(\alpha_i) = 1/4$ for any value of $\alpha_i > 0$.

The following result explicitly relates partner selection and item selection, again assuming a 2PL IRF.

Proposition 7. Under the same conditions as Proposition 6, let $\Delta^*(\alpha_i) = D$ for some constant $D \in (0, 1/4]$. Then

$$\alpha_i = \frac{2}{\delta} \ln \frac{1 - \sqrt{D}}{\sqrt{D}}.$$

The proposition shows that, for any desired level of item (non-) redundancy, D, the item discrimination must be chosen to be inversely proportional to the difference between the ability levels of the respondents. This implies that it will not be possible to select items that are both highly discriminating (i.e., strongly related to the performance domain) and also non-redundant, when group members have disparate levels of ability.

5.2. Summary

This section has considered how to design group assessments so as to avoid equivalence among four special cases of the RSC model. In particular, we addressed the conditions under which the expected performance of a dyad will be identical to that of either of its members, and referred to this situation in terms of (item or partner) redundancy. In order to avoid partner redundancy, partners should be chosen to have proximate levels of ability. In order to avoid item redundancy under a 2PL IRT model, items should be chosen to have difficulty equal to the average of the dyad's ability levels, and be moderately discriminating. We also showed that it will not be possible to find non-redundant items that are strongly related to the performance domain, when team members have disparate levels of ability. Interpreted more generally, this last result could be taken to suggest something along the lines of Vygotsky's (1978) zone of proximal development as a feasibility condition for designing group assessments based on social combination theory.

Two additional points should be mentioned. First, the conditions under which a group assessment will be non-redundant are not also conducive to distinguishing which of two individuals were contributing to a group's performance (i.e., distinguishing the Min and the Max models). This exemplifies the more general point that different design considerations will come into play depending on the intended purpose of a group assessment. Our stated purpose has been to distinguish the expected performance of a group from that of its individual members. Second, these results are all predicated on the RSC model, and in particular on the interpretation of the Independence model and the Additive model as lower and upper bounds on group performance, respectively. It will be interesting to consider extensions of these results to more general models.

6. Estimating the RSC model

The four models of group work that we have focussed on thus far are point hypotheses, in the sense that they do not have parameters to be estimated. To provide a more flexible approach, it would be desirable to instead infer the values of $a, b \in [0, 1]$ in the RSC model from group assessment data, leading to the following group IRF:

$$R_i = a\, P_{i1}\, Q_{i2} + b\, Q_{i1}\, P_{i2} + P_{i1}\, P_{i2}. \tag{13}$$

Identification of the weight parameters depends on the values of θ_1 and θ_2. In particular, if $\theta_1 = \theta_2$, then $P_{i1}\, Q_{i2} = Q_{i1}\, P_{i2}$ for all i, so that $a + b$ is identified but the individual weights are not. On the other hand, Propositions 3 and 4 show that $P_{i1}\, Q_{i2} \to 0$ as $\delta = \theta_2 - \theta_1 \to \infty$, in which case b is identified but a is not.

To avoid these cases of model unidentification, we instead focus on a "one-parameter" RSC model obtained by setting $a = b = w \in [0, 1]$. Using a single weight, the resulting IRF may be written as a linear interpolation between the lower and upper bounds of the RSC model,

$$R_i = w\, R_i^{11} + (1 - w)\, R_i^{00}. \tag{14}$$

The parameter w is then directly interpretable in terms of Steiner's (1972) concept of process loss, with $w = 0$ denoting complete process loss, and $w = 1$ denoting maximal group performance. When $w = 1/2$ we have the average of the two individuals' performance, $R_i = (P_{i1} + P_{i2})/2$, and when $\theta_1 \approx \theta_2$ this means that a group performs at about the same level as its individual members. Thus, when partners are matched on ability, values of $w < 1/2$ correspond to worse-than-individual performance, and values of $w > 1/2$ correspond to better-than-individual performance. The model also remains interpretable when $\Delta_{i12} = 0$, in which case it reduces to a weighted average of the performance of the two individuals, $R_i = w\, P_{i1} + (1 - w)\, P_{i2}$.

The item information function for the one-parameter RSC model is readily obtained as:

$$\mathcal{I}_i(w) = E\left[\frac{\partial^2}{\partial w^2} \ln f_i(Y_i \mid \boldsymbol{\theta})\right] = \frac{(P_{i1}\, Q_{i2} + Q_{i1}\, P_{i2})^2}{R_i(1 - R_i)}, \tag{15}$$

where f_i denotes the discrete mass function of Y_i. When partners are chosen such that $\theta_1 = \theta_2 = \theta$, the item information reduces to

$$\mathcal{I}_i(w) = [2\, w\, (1 - w) + w\, Q_i/P_i + (1 - w)\, P_i/Q_i + 1/2]^{-1}. \tag{16}$$

For $w \in (0, 1)$, this approaches zero when either P_i or Q_i approach zero. If, in addition, items are chosen such that $P_i = 1/2$, then Equation (16) further reduces to

$$\mathcal{I}_i(w) = [2\,w\,(1-w) + 3/2]^{-1}. \tag{17}$$

This last expression gives the item information when $\Delta_{i12} = 1/4$ (i.e., when item redundancy is minimized). We have not found analysis of $\mathcal{I}_i(w)$ under more general conditions to yield intuitive conclusions.

6.1. Estimation when θ_1 and θ_2 are Unknown

Up until now we have treated θ_1 and θ_2 as known parameters, thereby ignoring measurement error in the estimates of individual team members' abilities. In practice, this is not a realistic assumption. One way to address this situation is to estimate the parameter vector $\boldsymbol{u} = (\theta_1, \theta_2, w)$ using data from both an individual and group assessment simultaneously. We take this approach in our empirical example, and the maximum likelihood (ML) equations for obtaining estimates $\hat{\boldsymbol{u}}$ are given in the Appendix.

A second practical issue arises when the true parameter value $w \in \{0, 1\}$. When w is not in the interior of the parameter space, the usual asymptotic results for ML do not apply. Additionally, in practice we have found that values of $\hat{w} \in \{0, 1\}$ often arise due to large standard errors of θ_1, θ_2, or w. We illustrate this scenario in our simulation study.

To address this situation, we also consider modal a' posteriori (MAP) estimation of $\boldsymbol{u}$, with the estimating equations also presented in the Appendix. For the ability estimates we use a standard normal prior. For the weights we use a two-parameter Beta prior, with both parameters equal to $1 + \epsilon$ where ϵ is a small positive number. This results in a relatively flat prior over the range $[.05, .95]$, but a sharp decrease to zero for values approaching $\{0,\ 1\}$. An alternative would be a weakly informative Gaussian prior on a logit parameterization of w, say $w(a) = \exp(a)/[1 + \exp(a)]$. Centering $\mu_a = 0$ gives $w(\mu_a) = 1/2$, which, as discussed above, corresponds to the average of the two individuals' IRFs. This prior would be especially suitable when the individuals have proximate ability levels. However, transforming the logit back to the probability scale, the delta method yields the standard error (SE)

$$\text{SE}(\hat{a}) = w(\hat{a})\,[1 - w(\hat{a})] \times \text{SE}(\hat{a}),$$

which is excessively optimistic for values of $w(\hat{a})$ close to 0 or 1. Hence this prior is not well suited to addressing boundary values of $\hat{w}$.

Other estimation methods are available. For instance, expected a' posterior in three dimensions remains tractable by numerical integration. A "fully" Bayesian approach could be used to incorporate sampling uncertainty in the item parameters. Plausible values might be used when item-level data on individual assessments are not readily available. We leave developments along these lines to future research.

7. Simulation Study

This purpose of this small simulation is to illustrate the parameter recovery of the RSC model using the estimating equations presented in the Appendix. We consider both ML and MAP estimation, and illustrate the benefits of the latter when the individual assessment or the group assessment has few items. Readers interested in replicating our results or conducting further simulations can use the `R` package `cirt` and accompanying documentation, available at `github.com/peterhalpin/cirt`.

The simulation used 2PL IRFs for the items, and the following data-generating parameters:

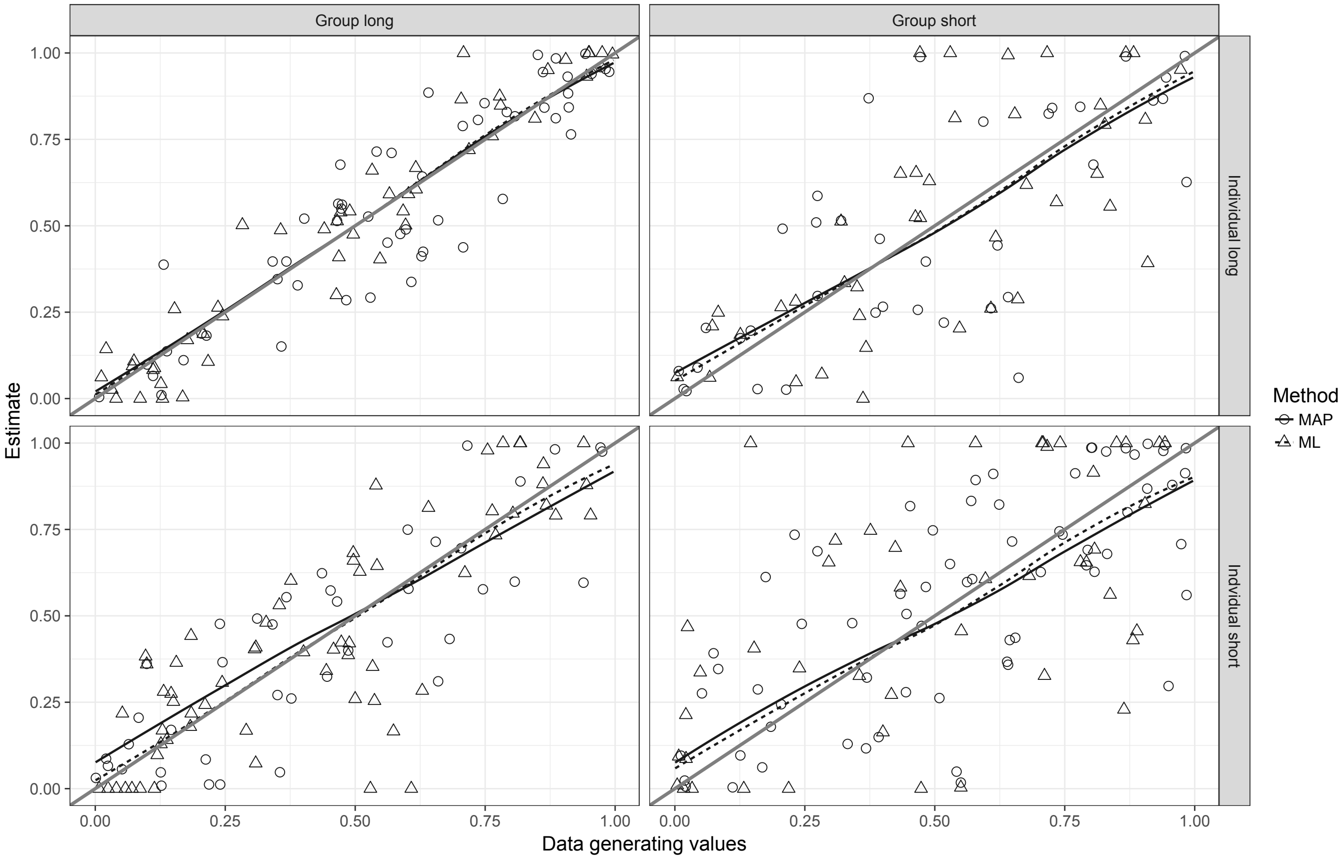
$$\beta_i \sim N(0, 1.3); \quad \alpha_i \sim \text{Uniform}(.6, 2.5) \quad \theta_j \sim N(0, 1); \quad w_k \sim \text{Beta}(1.05, 1.05)$$

The range of item parameters was chosen to realistically reflect the empirical example described in the following section. The total number of items generated was $I = 200$ with half used on the individual assessment and half on the group assessment. To simulate shorter assessments, a single random sample of 20 items was selected from each assessment; this shorter test length also corresponds to the empirical example described below. Data were simulated for $J = 1000$ respondents, which were matched at random into $K = J/2$ non-overlapping pairs. Parameter recovery was examined in each of four conditions obtained by crossing the long (100 items) and short (20 items) test forms of the individual and group assessment.

Figure 1 summarizes the parameter recovery of the two estimators in each of the four test length conditions. The grey line along the diagonal represents perfect agreement between the estimates and the data-generating values, and the dashed and solid lines represent the loess-smoothed ML and MAP estimates, respectively. The main take-aways are (a) both estimators perform well with a lot of items; but (b) when either or both of the assessments involved a small number of items, bias was apparent for both estimators; and (c), as expected, the bias tended to be larger for the MAP estimator, especially closer to the boundaries of $\{0, 1\}$. Also note that ML estimates (triangles) were observed on the boundaries, and the proportion of estimates on the boundary was larger for the shorter test lengths. This was not the case for the MAP estimates (circles).

Figure 2 displays the SEs for the each of the two estimators in each of the four conditions.

Figure 1: Parameter recovery of w for ML and MAP estimators, in the four test length conditions. The "long" condition denotes the full pool of 100 items and the "short" condition denotes a random subsample of 20 items, for each of the group and individual assessments. The dashed and solid lines represent the loess-smoothed ML and MAP estimates, respectively. The triangles and circles represent variation in the ML and MAP estimates, respectively, using a random sample of points from each condition.



The SEs were computed by inverting the expectation of the Hessian at the estimated values, as per the simultaneous estimation procedure described in the Appendix. Unsurprisingly, the standard errors were smaller when the group test was longer, and, for the short group test, inference about the performance of many dyads was highly unreliable. Also as expected, the standard errors of the MAP estimates were substantially lower for values close to the boundaries of $\{0, 1\}$.

Comparing the MAP estimates close to the boundary values in Figures 1 and 2, the bias-variance trade off induced by the prior distribution is apparent. However, for parameter values in the range $[.25, .75]$, the estimates were very similar.

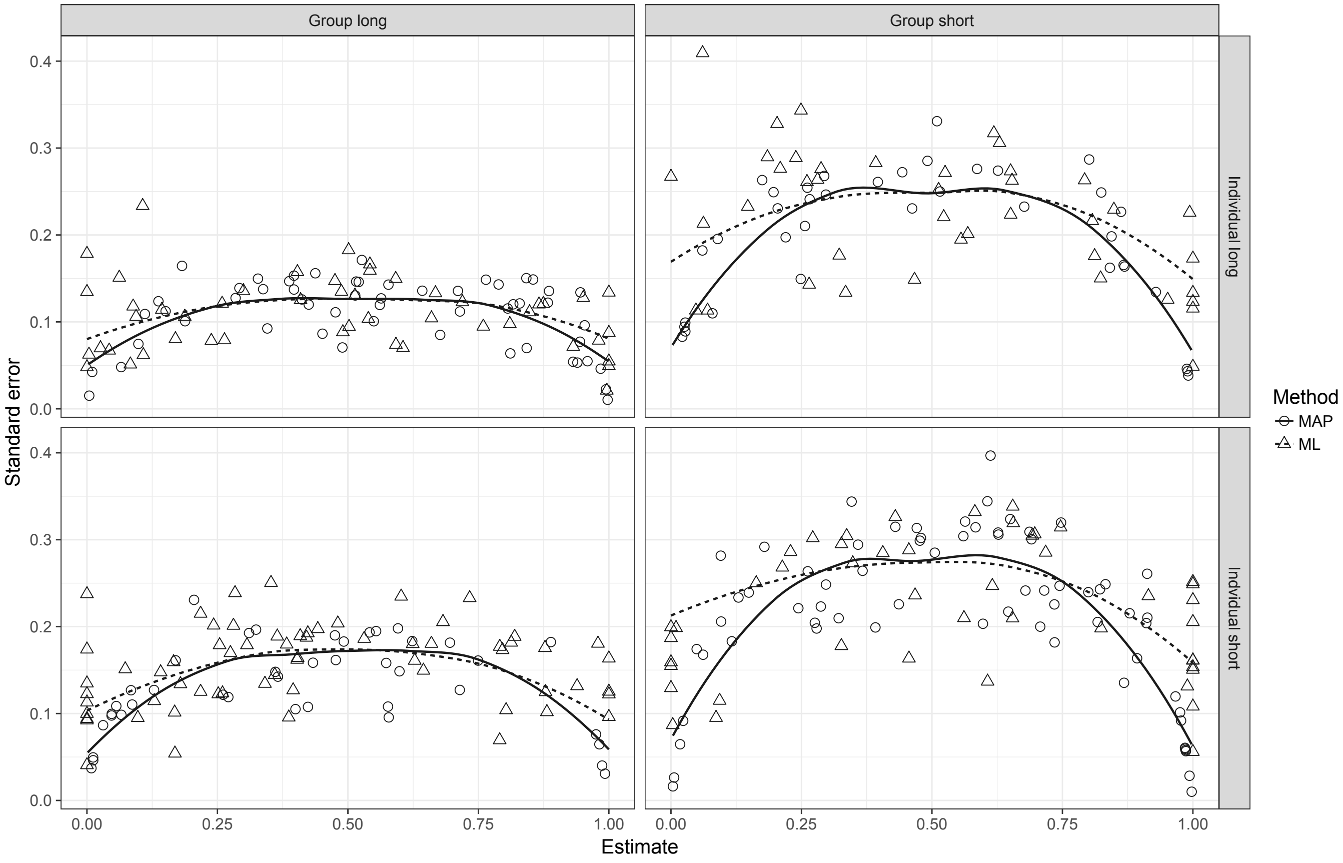
8. Empirical example

In this section we apply the foregoing results to address the following focal question: is there evidence of better-than-individual group performance when real dyads work together? As preliminary analyses, we also investigated (a) the measurement invariance of items calibrated for individual performance, when used in a group setting, and (b) the goodness of fit of the one-parameter RSC model to data collected from real dyads.

8.1. Sample and Procedure

Respondents were solicited using Amazon Mechanical Turk (AMT). Approximately 5000 AMT workers were pre-screened using a demographic survey, and these pre-screened workers constituted the sampling frame for the present study. The sampling frame was comprised exclusively of AMT workers who self-reported to live in the United States and to speak English as

Figure 2: Standard errors of the ML and MAP estimates, in the four test length conditions. The "long" condition denotes the full pool of 100 items and the "short" condition denotes a random subsample of 20 items, for each of the group and individual assessments. The dashed and solid lines represent the loess smoothed ML and MAP standard errors (SEs), respectively. The triangles and circles represent variation in the ML and MAP SEs, respectively, using a random sample points from each condition. The SEs were computed via the expectation of the Hessian.



their first language. The median age was 32 years, with an interquartile range of [27, 40]. The majority of the sampling frame (71%) self-identified as being of "White" ethnicity, 51% reported being female, and 88% reported having at least one year of post-secondary education. Two independent samples were taken from the sampling frame, a calibration sample ($N = 528$) and a research sample ($N = 322$).

The calibration sample was used to estimate item parameters of the 2PL model for a pool of $I = 60$ twelfth-grade mathematics items obtained from previous administrations of the National Assessment of Educational Progress (NAEP). The mathematical content of the items was preserved, but they were modified to be delivered online and to use numeric response rather than multiple-choice formatting. Additionally, participants were instructed to complete the assessment in whatever conditions they deemed suitable, and were explicitly permitted to use a calculator and the internet. It was not possible to enforce any other requirements, so these were de facto testing conditions of the present study. Item parameters of the 2PL model were estimated using maximum likelihood and a total of three items were removed from the item pool due to poor item fit (non-montone IRFs). The remaining items had parameter estimates in the following ranges: $\hat{\beta}_i \in [-3.80, 2.62]$ and $\hat{\alpha}_i \in [0.65, 2.86]$.

In the research sample, all respondents were assessed under both individual and group testing conditions, and the content of the individual and group forms was counterbalanced. In the individual testing condition, respondents were administered a form consisting of 20 items from the calibration sample, which was used to estimate their mathematical ability. After completing the individual assessment, respondents were routed to a second form consisting of another 20 items from the calibration sample. Before commencing the second form, respondents were provided with same instructions as for the individual form, with the exception that (a) they would be paired with an anonymous partner, and (b) they were encouraged to work with their partner to ensure that both individuals arrived at the correct response. After acknowledging the instructions, respondents were randomly paired based on their arrival in the routing queue, and they interacted with their partner via online chat.

The online testing platform led to two main limitations in the study design. First, items could not be adaptively administered, or even randomized within forms. Second, it was originally intended to match respondents based on their performance on the individual pre-test, but it proved to be infeasible to implement anything other than matching based on arrival times. Thus, the order of the individual and group testing conditions was not counterbalanced, and we were unable to make use of information from the individual test when selecting partners for the group test. Despite these limitations, we are not aware of any other dataset that provides an opportunity to study small groups working together on calibrated test items.

8.2. Measurement invariance

We assessed measurement invariance using the calibration sample and the individual (not conjunctively-scored) response patterns of participants in the group testing condition. This resulted in an independent samples design, where respondents in the group testing condition were nested within dyads. It is important to note that, due to the counterbalancing of the individual and group tests, each item in the group testing condition was responded to by only half (161) of the participants in the research sample. Therefore the following results should be regarded as highly preliminary.

The analysis was implemented in Mplus 7 (L. K. Muthén & Muthén, 2015) using the cluster-robust maximum likelihood estimator (B. O. Muthén & Satorra, 1995). Measurement invariance was assessed using the Satorra-Bentler adjusted likelihood ratio test (Satorra & Bentler, 2010) of the scalar and metric models against the configural model. Respondents in the group testing condition were clustered within dyads, and respondents in the calibration sample each formed their own cluster of one.

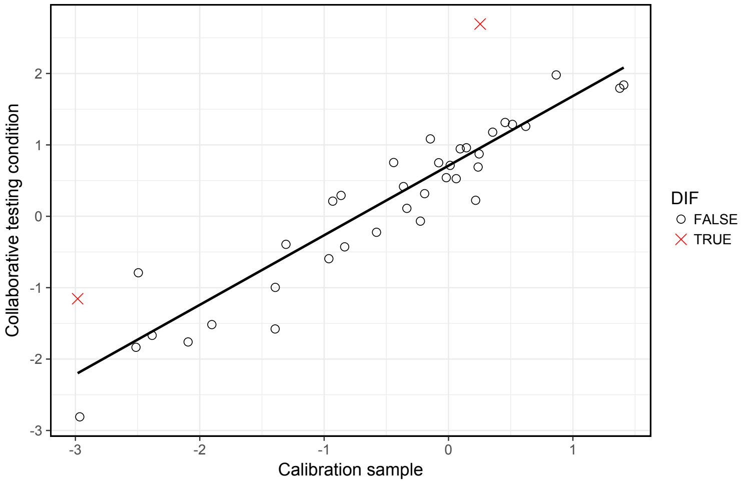
The results are summarized in Table 1. The scalar model, but not the metric model, had worse fit than the configural model. The estimated difficulty parameters in the configural model are plotted in Figure 3. After removing the two items indicated in the figure, measurement invariance was again assessed. With the two items removed, the scalar model fit reasonably well, as indicated in the last row of Table 1. All subsequent analyses omitted these two items from the group testing condition.

Table 1: Measurement invariance in individual and group testing conditions

Model	LR	*df*	*p*
Metric	36.971	37	.470
Scalar	101.071	74	.020
Scalar (w/ drop)	84.358	70	.116

Note: LR denotes the Satorra-Bentler adjusted likelihood ratio test against the configural model, *df* its degrees of freedom, and *p* its right-tail probability. ‘Scalar (w/ drop)’ denotes the scalar invariance model after dropping the two items indicated in Figure 3.

Figure 3: Item difficulty estimates for the calibration sample and the group testing condition, in the configural model. Reference line obtained form ordinary least squares regression, with slope = .975 and intercept = .701.



The results of the measurement invariance analysis suggest that individual and group performance on most of the mathematics items were indeed commensurable. Mathematics ability was higher in the group testing condition with a standardized group mean difference of $d = 0.598$ (SE = 0.162). It may be concluded that, on average, respondents performed better when working in dyads than working individually. However, it is important to note that this analysis does not tell us whether any group exhibited better-than-individual performance.

8.3. Goodness of fit

We assessed goodness of fit of the RSC model to the conjunctively-scored group response pattern of each dyad using a parametric bootstrap. The RSC model was fitted to the $N = 161$ conjunctively-scored response patterns using the MAP estimator. For each dyad, $R = 500$

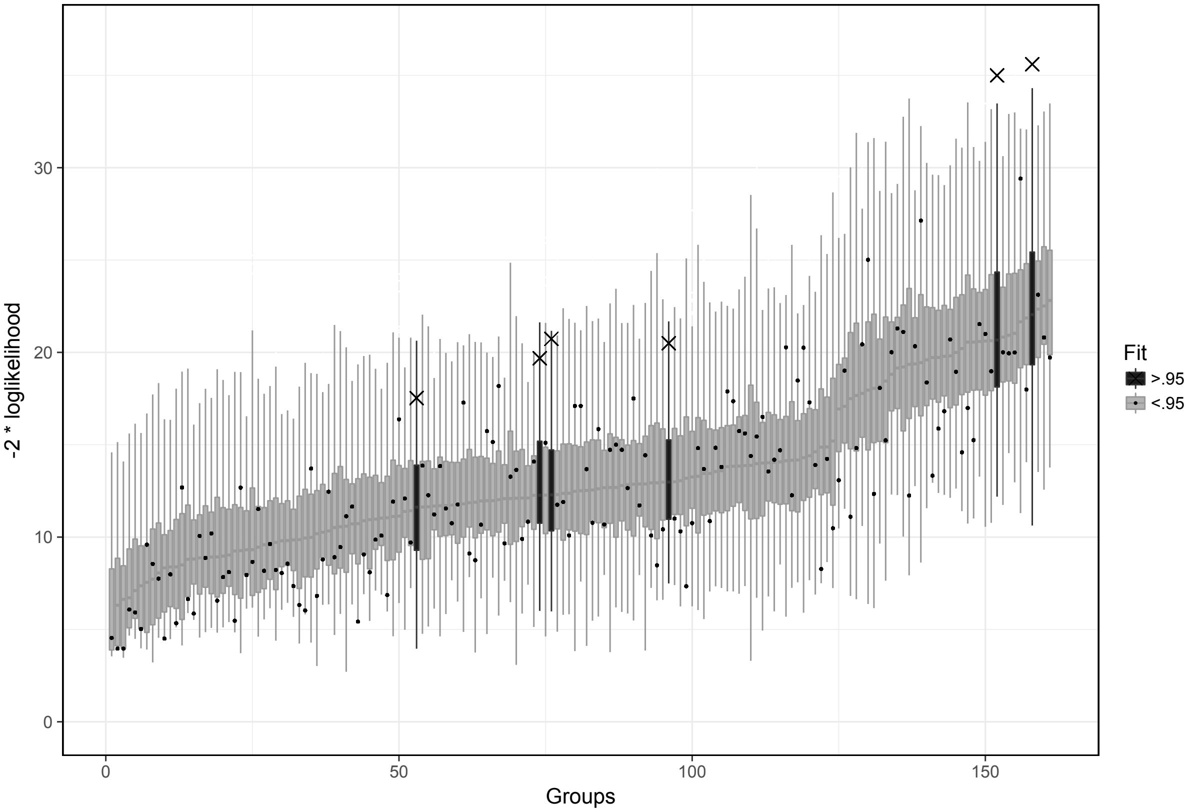
replicated responses to the group assessment were generated from the fitted model, treating the estimated values as the data-generating parameters. The weight of the RSC model was re-estimated for each generated data set, again using the MAP estimator but treating the data-generating values of θ as known. The log-likehood was computed for each of the re-estimated models, yielding a bootstrapped sampling distribution for the log-likelihood under the assumption that the fitted model was the data-generating distribution. We then compared the observed value of the log-likelihood to the bootstrapped sampling distribution. This approach is somewhat similar to Levine and Rubin's (1979) person fit statistic, except applied to the log-likelihood of the RSC model, rather than a more standard IRT model.

The results are summarized in Figure 4. Groups whose fit would be rejected at the 5% (one-tailed) significance level are indicated. A total of six groups (3.7%) would be rejected using this criterion, which corresponds reasonably well to the nominal rejection rate. We conclude that RSC model adequately represented the performance of the groups in the present sample.

8.4. Results

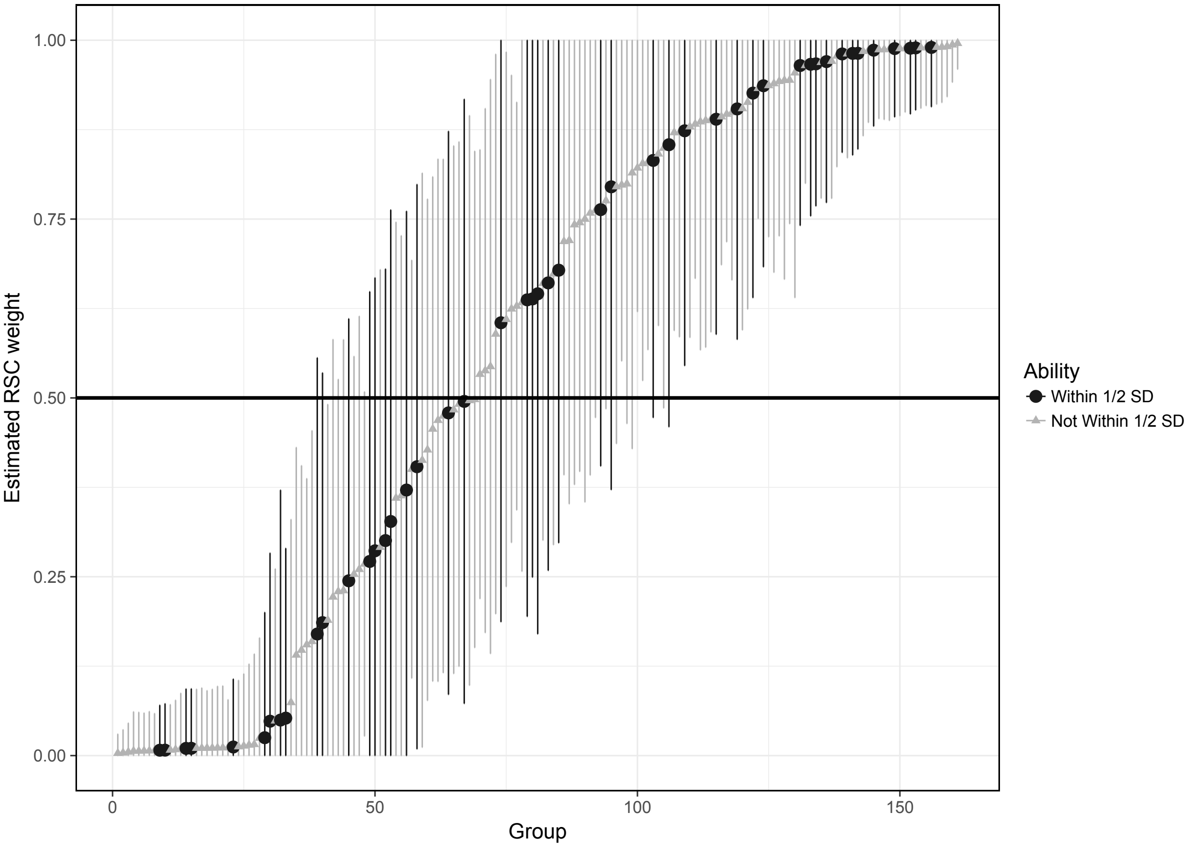
Finally we consider the MAP estimates of w, denoted $\hat{w}$, for the observed data. The results are summarized in Figure 5. As expected from the data simulation, inference about w was highly unreliable for most groups, due to the short length of both test forms. Therefore, we simply interpret whether the approximate 95% confidence (credible) interval on $\hat{w}$ included the value of 1/2. Recall that $w = 1/2$ corresponds to the average of the expected performance of the two group members working individually. Hence, for partners with proximate levels of ability, $w >> 1/2$ is evidence of better-than-individual performance. In the figure, proximate ability was

Figure 4: Goodness of fit for each group. Reference distributions for log-likelihood of the one-parameter RSC model were generated using 500 replications for each dyad. Groups denoted by black crosses and boxplots had fitted log-likelihoods which were improbable under the assumption that the fitted model was the data-generating distribution.



operationalized as being within 1/2 standard deviation unit on the ability scale. This value was chosen because it is easily interpretable and corresponded to about $2 \times \mathrm{SE}\,(\hat{\theta})$ for most respondents. From the figure we can see that a total of 17 out of 47 (36 %) dyads who were matched on ability were also inferred to have performed at a higher level than either respondent working independently. An additional 9 groups (19 %) were inferred to have performed at a level lower than either respondent working independently.

Figure 5: Estimated weights with approximate 95% confidence (credible) intervals for each group. Confidence intervals were computed using a Gaussian reference distribution with the approximate posterior standard deviation of the RSC weight, computed via the expected value of the Hessian. Intervals that included the boundary values of $\{0, 1\}$ were truncated. Groups denoted by black circles and error bars had partners whose ability estimates were within 1/2 standard deviation unit of each other.



9. Conclusions

This paper has shown how the social combination theory of group problem solving can be used to extend existing psychometric models to collaborative settings. In particular, we have focussed on a restricted social combination (RSC) model for pairwise group work under a conjunctive-scoring rule for binary (correct / incorrect) responses. The model is restricted in the sense that it imposes non-trivial lower and upper bounds on groups' expected performance. The lower bound requires that groups do not perform worse than their individual members working independently. The upper bound was motivated by empirical and theoretical research related to Lorge & Solomon's (1955) Model A, and in particular by Steiner's (1972) discussion of process loss for intellective tasks. The development of more general models that allow for the plausibility of these bounds to be tested against group assessment data is a clear priority for future research. However, the RSC model provides a starting place for this research, and our re-formulation of Davis' (1973) general social combination model in Equation (6) provides a framework for moving forward.

The RSC model was shown to yield latent monotonic group IRFs (Proposition 1) and to have a number of relatively intuitive implications for the design of group assessments. In particular, we outlined conditions under which the expected performance of a dyad will be identical to that of either of its members (Proposition 2), and referred to this situation in terms of redundancy. In Propositions 3 through 6, we showed that redundancy can be avoided via team composition (selection of partners with proximate levels of ability) and item selection (for the 2PL, moderately discriminating items targeted between the ability levels of the two partners). We also showed that it will not be possible to find non-redundant items that are strongly related to the performance domain (i.e., discriminating), when team members have disparate levels of ability (Proposition 7). It is important to keep in mind that these results are predicated on the RSC model. The derivation of optimal design conditions for group assessments under more general models is another clear priority for future research.

Because the RSC model is not identified under all team composition conditions, we proposed the one-parameter RSC model as a viable alternative for data analysis and inference. The parameter of this model interpolates between the lower and upper bounds of the RSC model, and has a direct interpretation in terms of process loss – a value of zero indicates complete process loss, and a value of one indicates optimal group performance, and a value of 1/2 corresponds to the average performance of the two group members. When partners have proximate levels of ability, values disparate from 1/2 can also be interpreted in terms of better- (or worse-) than-individual performance. We provided equations for maximum likelihood and modal a' posteriori estimation of the one-parameter RSC model (see Appendix), and used data simulation to illustrate the advantages of the latter with short tests. In our real data example, we provided a preliminary evaluation of the measurement invariance of item parameters under individual and group testing conditions, considered the goodness of fit the one-parameter RSC model to data from real dyads, and concluded that a substantial number of dyads did indeed demonstrate better-than-individual performance when working together on an online mathematics test. Unfortunately, the online testing platform used in the real data example did not allow for the results on team composition and item selection to be put into action. Perhaps the highest priority for future research is the design and implementation of software for delivering group assessments.

In addition to those already mentioned, there are number of future directions that can advance psychometric research on group assessments. Some relatively obvious extensions include models for (a) non-binary responses, (b) groups with more than two members, and (c) different types of scoring rules. The development of psychometric models for group performance can also be facilitated by the design of group tasks that require respondents to exhibit the types of skills theorized to support productive group work. A number of innovative group tasks have already been implemented in large scale testing platforms (Griffin & Care, 2015; Organisation for Economic Co-operation and Development, 2013), with a recent white paper outlining the implementation landscape for NAEP (Fiore et al., 2017). Finally, we suggest that models and software that support multiple group memberships will be a major technical challenge to be addressed before group assessments are ready for 'prime time.' Our initial work along these lines suggests that multiple group memberships can be used to identify the two-parameter RSC model, leading to inference about individual respondents' performance in group settings. Given ongoing progress in these areas, we hope that group assessments will be a practical reality in the near future.

10. Appendix

10.1. Proofs

This section contains the proofs for Propositions 1, and 3 through 7. We let $j = 1, 2$ denote the members of an arbitrary dyad, and assume that $\theta_1 \leq \theta_2$ by choice of notation. Subscripts for items are omitted. Several proofs require derivatives of monotonic functions, which the reader will recall are defined almost everywhere on their domain.

10.1.1. Proof of Proposition 1

Let $f, g : \mathbb{R} \to [0, 1]$ be monotone non-decreasing functions and let $a, b \in [0, 1]$ be fixed constants. The function

$$\begin{aligned} h(x, y) &= a\, f(x)[1 - g(y)] + b\, [1 - f(x)]g(y) + f(x)g(y) \\ &= a\, f(x) + b\, g(y) + (1 - a - b)\, f(x)g(y) \end{aligned} \tag{18}$$

is seen to be non-decreasing in x for fixed y by considering it partial derivative in x and noting that $df/dx = f'(x) \geq 0$:

$$\begin{aligned} \frac{\partial}{\partial x} h(x, y) &= a\, f'(x) + (1 - a - b)\, f'(x)\, g(y) \\ &= a\, f'(x)\, [1 - g(y)] + (1 - b)\, f'(x)\, g(y) \geq 0. \end{aligned} \tag{19}$$

A similar argument shows that Equation (18) is also non-decreasing in y, and Proposition 1 follows directly.

10.1.2. Proof of Proposition 3

Let $f(x, y) = x(1 - y)$ with $0 < x \leq y < 1$. We show f is strictly concave with global maximum $f(1/2, 1/2) = 1/4$.

A sufficient condition for f to be strictly concave is that $\boldsymbol{u}' H \, \boldsymbol{u} < 0$, where $H = \left(\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right)$ is the Hessian of f and $\boldsymbol{u} = (u_1, u_2)$ is in the domain of f. The quadratic form reduces to $q = -2\, u_1 u_2$, and the u_i are strictly positive, so $q < 0$.

The global maximum can be found by applying the Karush-Kuhn-Tucker (KKT) conditions for constrained optimization as follows (see e.g., Boyd & Vandenberghe, 2004). The only inequality that is active at the proposed solution is $g(x, y) = x - y \leq 0$, so the objective function and its gradient may be written, respectively, as

$$L(x, y, \mu) = f(x, y) - \mu \, g(x, y) = x(1 - y) - \mu(x - y),$$

$$\nabla L(x, y, \mu) = \begin{bmatrix} 1 - y + \mu \\ -x - \mu \end{bmatrix}.$$

The KKT conditions state that any local maximum (x^*, y^*) of f must satisfy $\nabla L(x^*, y^*, \mu) = \mathbf{0}$, and $\mu \, g(x^*, y^*) = 0$ for $\mu \neq 0$. These equations are readily solved to show $y^* = x^* = 1/2$.

10.1.3. Proof of Proposition 4

Part 1 of the proposition follows from directly from the definition of θ_0 and the global maximum of $\Delta(P_1, P_2)$ derived in Proposition 3.

Part 2 additionally uses the result (from Proposition 3) that $\Delta(P_1, P_2)$ is strictly concave, and the assumption (from Proposition 4) that $P(\theta)$ is strictly increasing on $\mathcal{N}$, which together imply that $\Delta(u_{12})$ is strictly decreasing in each coordinate of $u_{12} = (\theta_0 - \theta_1, \theta_2 - \theta_0)$, for $\theta_1, \theta_2 \in \mathcal{N}$. The result then follows from writing $\delta = (\theta_2 - \theta_0) + (\theta_0 - \theta_1)$.

10.1.4. Proof of Proposition 5

Let $P(z_j) = [1 + \exp\{-z_j\}]^{-1}$ with $z_j = \alpha(\theta_j - \beta)$ and $Q(z_j) = 1 - P(z_j)$. We show that

$$\arg\max_{\beta} \{P(z_1)\, Q(z_2)\} = (\theta_1 + \theta_2)/2.$$

First note that

$$\frac{\partial}{\partial\beta} P(z_1)\, Q(z_2) = \alpha\, P(z_1)\, Q(z_2)\; [P(z_2) - Q(z_1)].$$

Setting this to zero gives

$$Q(z_1) = P(z_2) \quad \Leftrightarrow \quad P(-z_1) = P(z_2) \quad \Leftrightarrow \quad -z_1 = z_2, \tag{20}$$

hence there is a single critical point at $\beta^* = (\theta_1 + \theta_2)/2$. To show that this is a local maximum, we first find the second derivative,

$$\frac{\partial^2}{\partial\beta} P(z_1)\, Q(z_2) = \alpha^2\, P(z_1)\, Q(z_2)\, \left([P(z_2) - Q(z_1)]^2 - P(z_1)\, Q(z_1) - P(z_2)\, Q(z_2)\right),$$

then use Expression (20) to write

$$\begin{aligned} \frac{\partial^2}{\partial\beta} P(z_1)\, Q(z_2)\bigg|_{\beta^*} &= \alpha^2\, P(z_1)^2\, \left([Q(z_1) - Q(z_1)]^2 - 2P(z_1)\, Q(z_1)\right) \\ &= -2\, \alpha^2\, P(z_1)^3\, Q(z_1) < 0. \end{aligned}$$

Since there is only a single critical point and this is a local maximum, it follows that β^* must also be the global maximum that $P(z_1)\, Q(z_2)$ is strictly concave in β.

10.1.5. Proof of Proposition 6

Using the same notation as above, let $z_j^* = \alpha(\theta_j - \beta^*)$. We show that $P(z_1^*)\,Q(z_2^*)$ is monotone non-increasing in α as follows:

$$\frac{\partial}{\partial\alpha} P(z_1^*)\,Q(z_2^*) = \frac{\partial}{\partial\alpha}[P(z_1^*)]^2 = 2\,(\theta_1 - \beta^*)\,P(z_1^*)\,Q(z_1^*) \leq 0.$$

The first equality uses Expression (20), and the inequality follows since $\theta_1 \leq \beta^*$ by choice of subscripts $j = 1, 2$.

10.1.6. Proof of Proposition 7

Using the same notation as above, the proposition follows by using the following equalities to solve for α

$$D = \Delta^*(\theta_{12}) = P(z_1^*)\,Q(z_2^*) = [P(z_1^*)]^2.$$

10.2. Estimating equations

This section provides the necessary derivatives for simultaneously estimating the parameter vector $\boldsymbol{u} = (\theta_1, \theta_2, w)$ for an arbitrary dyad. We consider both maximum likelihood (ML) estimation and the modal a'posteriori (MAP) method described in the main paper. Referring to Equations (1) through (3), let θ_r and $\boldsymbol{X}_r$ denote the latent trait and response pattern, respectively, for respondent r. The group response vector from Equations (4) and (5) is denoted $\boldsymbol{Y}$ and the parameter w is the weight from the one-parameter RSC model in Equation (14). We let $P_{ir} = P_i(\theta_r)$ denote the IRF for item i on an individual assessment, and $R_j = R_j(\boldsymbol{u})$ denote the group IRF for item j on a group assessment. Estimation using the equations outline in this section is implemented in the `R` package `cirt` available at `github.com/peterhalpin/cirt`.

Using the local independence assumptions for individual and group assessments, the log-likelihood of interest is

$$\ell(\boldsymbol{u} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}) = \sum_i \ell(\theta_1 \mid X_{i1}) + \sum_i \ell(\theta_2 \mid X_{i2}) + \sum_j \ell(\boldsymbol{u} \mid Y_j) \tag{21}$$

where

$$\ell(\theta_r \mid X_{r1}) = x_{ir} \, \ln(P_{ir}) + (1 - x_{ir}) \, \ln(1 - P_{ir})$$

and

$$\ell(\boldsymbol{u} \mid Y_j) = y_j \, \ln(R_j) + (1 - y_j) \, \ln(1 - R_j).$$

Methods for estimating θ_r via $\ell(\theta_r \mid X_{r1})$ are well known (e.g., Baker & Kim, 2004), so we focus on estimation of w via $\ell = \ell(\boldsymbol{u} \mid Y_j)$.

Its gradient is

$$\nabla \ell = \frac{\partial}{\partial \boldsymbol{u}} \ell = \sum_j m_j \left[\frac{\partial}{\partial \theta_1} R_j \quad \frac{\partial}{\partial \theta_2} R_j \quad \frac{\partial}{\partial w} R_j \right]^T \tag{22}$$

where

$$m_j = \frac{y_j}{R_i} - \frac{1 - y_j}{1 - R_j}.$$

Letting $P'_{ir} = \frac{\partial}{\partial \theta_r} P_{ir}$ denote the derivatives of the individual IRFs, the derivatives of the group IRFs in Equation (14) can be written

$$\frac{\partial}{\partial \theta_r} R_j = [w + (1 - 2w) \, P_{js}] \, P'_{ir}$$

and

$$\frac{\partial}{\partial w} R_j = P_{jr} + P_{js} - 2P_{jr} \, P_{js} = P_{jr} Q_{js} + Q_{jr} P_{js}.$$

Letting $H(\ell) = \{h_{rs}\}$ denote the Hessian of ℓ, its elements are given by

$$h_{rs} = \frac{\partial^2}{\partial u_r \partial u_s}\ell = m_j \, \frac{\partial^2}{\partial u_r \partial u_s} R_j - n_j \, \frac{\partial}{\partial u_r} R_j \, \frac{\partial}{\partial u_s} R_j \qquad r, s = 1, 2, 3 \tag{23}$$

with

$$n_j = \frac{y_j}{R_j^2} + \frac{1 - y_j}{(1 - R_j)^2}.$$

Letting $P''_{ir} = \frac{\partial^2}{\partial \theta_r^2} P_{ir}$, the necessary second derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial \theta_r^2} R_j &= [w + (1 - 2w)\, P_{js}]\, P''_{jr} \\
\frac{\partial^2}{\partial \theta_r \partial \theta_s} R_j &= (1 - 2w)\, P'_{jr}\, P'_{js} \\
\frac{\partial^2}{\partial \theta_r \partial w} R_j &= P'_{jr}\, (1 - 2P_{js}) \\
\frac{\partial^2}{\partial w^2} R_j &= 0.
\end{aligned}$$

ML estimation of w can then proceed using Equations (21) through (23) and the provided derivatives, with standard errors computed by inverting either the observed or expected Hessian. In the latter case, the terms m_j vanish under expectation, and the standard errors can be obtained using only the first-order derivatives of the individual and group IRFs.

When considering MAP rather then ML estimation, the likelihood in (21) is replaced by the posterior distribution of $\boldsymbol{u}$,

$$p(\boldsymbol{u} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y}) \propto p(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{Y} \mid \boldsymbol{u}) \times p(\boldsymbol{u}). \tag{24}$$

As described in the main paper, we assume that $p(\boldsymbol{u}) = \prod_k p(u_k)$ with $\theta_r \sim N(0, 1)$ and $w \sim \text{Beta}(u, v)$, where Beta denotes the two-parameter beta distribution and $u = v = 1 + \epsilon$ with ϵ

being a small positive number. Then

$$\ln p(\boldsymbol{u}) = -\frac{1}{2}(\theta_1^2 + \theta_2^2) + \epsilon \ln(w - w^2) + C \tag{25}$$

where C is a constant that does not depend on $\boldsymbol{u}$. MAP estimation of w proceeds by using

$$\nabla \ell + \frac{\partial}{\partial \boldsymbol{u}} \ln p(\boldsymbol{u}) \quad \text{and} \quad H(\ell) + \frac{\partial^2}{\partial \boldsymbol{u} \partial \boldsymbol{u}^T} \ln p(\boldsymbol{u})$$

in place of equations (22) and (23). The required first and second derivatives in θ_r are

$$\frac{\partial}{\partial \theta_r} \ln p(\boldsymbol{u}) = -\theta \quad \text{and} \quad \frac{\partial^2}{\partial \theta_r^2} \ln p(\boldsymbol{u}) = -1$$

and for the weights,

$$\frac{\partial}{\partial w} \ln p(\boldsymbol{u}) = \epsilon \frac{1 - 2w}{w - w^2} \quad \text{and} \quad \frac{\partial^2}{\partial w^2} \ln p(\boldsymbol{u}) = -\epsilon \left(\frac{2}{w - w^2} + \left(\frac{1 - 2w}{w - w^2} \right)^2 \right).$$

References

Aronson, E., Blaney, N., Stephan, C., Sikes, J., & Snapp, M. (1978). *The Jigsaw Classroom.* Beverly Hills, CA: Sage.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511804441

Cohen, E. G., Lotan, R. a., Abram, P. L., Scarloss, B. a., & Schultz, S. E. (2002). *Can Groups Learn?* (Vol. 104) (No. 6). doi: 10.1111/1467-9620.00196

Davis, J. H. (1973). Group decision and social interaction: A theory of social decision schemes. *Psychological Review*, *80*(3), 97–125. doi: 10.1037/h0021465

Davis, J. H. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950-1990. *Organizational Behavior and Human Decision Processes*, *52*(3-38).

Fiore, S. M., Graesser, A., Greiff, S., Griffin, P., Gong, B., Kyllonen, P., … von Davier, A. (2017). *Collaborative problem solving: Considerations for the National Assessment of Educational Progress* (Tech. Rep.). Washington, DC: National Center for Educational Statistics.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills: Methods and approach.* New York: Springer.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills.* New York: Springer.

Heckman, J. J., & Kautz, T. (2014). *Fostering and measuring skills: Interventions that improve Character and cognition* (Tech. Rep. No. 19656). doi: 10.1017/CBO9781107415324.004

Herman, J., & Hilton, M. (2017). *Supporting Students' College Success: The Role of Assessment of Intrapersonal and Interpersonal Competencies* (Tech. Rep.). Washington, DC: The National Academies Press. doi: 10.17226/24697

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional Association and Unidimensionality in Monotone Latent Variable Models. *The Annals of Statistics*, *14*(4), 1523–1543. doi: 10.1214/aos/1176350174

Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, *25*(3), 258–272. doi: 10.1177/01466210122032064

Laughlin, P. R. (1980). Socal combination processes of cooperattive, problem-sovling groups as verbal intellective tasks. In *Progress in social psychology* (pp. 127–155). Hillsdale, NJ: Erlbaum.

Laughlin, P. R. (2011). *Group Problem Solving.* Princeton, NJ: Princeton University Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269–290.

Lippman, L. H., Ryberg, R., Carney, R., & Moore, K. A. (2015). *Key "soft skills" that foster youth workforce success: Toward a consensus across fields. Child Trends Publication #201524* (Tech. Rep.). Washington, DC: Child Trends, Inc.

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf

(Eds.), *Handbook of research on computational tools for real-world skill development* (pp. 344–359). Hershey, PA: IGI-Global. doi: 10.1136/bmj.330.7485.0-h

Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. *Psychometrika*, *20*(2), 139–148. doi: 10.1007/BF02288986

McGrath, J. E. (1984). *Groups: Interaction and performance* (Prentice-Hall, Ed.). Englewood Cliffs, NJ.

Mesmer-Magnus, J. R., & DeChurch, L. (2009). Information sharing and team performance: A meta analysis. *Journal of Applied Psychology*, *94*(2), 525–546.

Muthén, B. O., & Satorra, A. (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, *25*, 267–316.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus 7 [computer software].* Los Angeles, CA: Muthén & Muthén.

National Research Council. (2011). *Assessing 21st Century Skills* (Tech. Rep.). Washington DC. doi: 10.17226/13215

Organisation for Economic Co-operation and Development. (2013). *PISA 2015 Draft Collaborative Problem Solving Framework* (Tech. Rep.).

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century.* doi: 0-309-25649-6

Poundstone, W. (1992). *Prisoner's Dilemma.* New York: Doublday.

Reckase, M. (2009). *Multdimensional Item Response Theory.* New York: Springer.

Salas, E., Cooke, N. J., & Rosen, M. a. (2008, jun). On Teams, Teamwork, and Team Performance: Discoveries and Developments. *Human Factors: The Journal of the Human*

Factors and Ergonomics Society, *50*(3), 540–547. doi: 10.1518/001872008X288457

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248.

Shiflett, S. (1979). Toward a general model of small group productivity. *Psychological Bulletin*, *86*(1), 67–79. doi: 10.1037/0033-2909.86.1.67

Smoke, W. H., & Zajonc, R. B. (1962). On reliability of group judgements and decisions. In J. H. Criswell, H. Solomon, & P. Suppes (Eds.), *Mathematical methods in small group processes* (pp. 322–333). Stanford, CA: Stanford University Press.

Stasser, G., & Titus, W. (2003, oct). Hidden Profiles: A Brief History. *Psychological Inquiry*, *14*(3-4), 304–313. doi: 10.1080/1047840X.2003.9682897

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring Hard-to-Measure Student Competencies* (Tech. Rep.). Santa Monica, CA: RAND Corporation.

Steiner, I. D. (1972). *Group Proceses and Productivity.* New York, NY: Academic Press.

Vgotsky, L. (1978). *Mind in Society.* Cambridge, MA: Harvard University Press.

von Davier, A., Kyllonen, P., & Zhu, M. (2017). *Innovative Assessments of Collaboration.* New York, NY: Spinger.

Vuong, Q. A. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrika*, *57*(2), 307–333.

Webb, N. M. (1995). Group Collaboration in Assessment: Multiple Objectives, Processes, and Outcomes. *Educational Evaluation and Policy Analysis*, *17*(2), 239–261. doi: 10.3102/01623737017002239