The CENTER for VICTIMS of TORTURE

Brief Ethnographic Interviewing Manual Appendix I: Statistical Analysis for Sort Method

Greg Vinson, Ph.D.
gvinson@cvt.org

Overview

The interview and sort method described in this document has traditionally been used to define and exemplify domains. It is a useful way to build theory, begin measure development, or investigate and define conceptual domain(s) or constructs (Anderson & Wilson, 1997; Bownas & Berdain, 1988; Flanagan, 1954). In essence, it is simply a systematic method to make sense of observations about some thing (Flanagan, 1954). It is a data-driven method for analyzing real world observations provided by others. In the cross-cultural context in which the Center for Victims of Torture (CVT) and many of our colleagues operate, it can be a way to make systematic observations about some thing using responses, sorters, and language of a different population. In this, it is less likely to impose other external perspectives, more likely to be sensible to the given population, and more likely to lend itself to the creation of reliable and valid measures of population-defined constructs (how CVT typically uses the method).

This method collects open-ended responses (here, called *incidents*) about some domain of interest and uses a card sort task to make sense of those responses. Many users of the method may opt to reach a final derivation of categories through some sort of group consensus. There are also several statistical options. Because all sorters sort the same incidents, it is possible to statistically model a final consensus model that incorporates all the sort information through quantitative data reduction techniques. The advantage of this method is that it is more apparently objective than coming to consensus through group processes and less prone to biases related to group processes. Also, all sorters and incidents are equally weighted in the analysis.

The essence of the statistical method is that a type of correlation matrix is created between all incidents, where incidents pairs that are more often sorted together are more highly correlated whereas those that are paired less frequently are less correlated. Following a specific procedure to calculate this matrix, it can be subjected to any number of conventional statistical data reduction techniques to determine the categorical structure of the incidents (e.g., Multi-Dimensional

Scaling, Principle Components Analysis, and Factor Analysis). Subsequent parts of this supplement will primarily concern the creation of the matrix and the subsequent analysis.

Such a method of data reduction for sort data is likely new to most. How does it compare to a group consensus approach? First, we are not aware of any formal research where the two methods are explicitly compared. This type of analysis is fairly novel. As applied, it has been rare when both approaches are used. It is possible that you could get qualitatively different results from both approaches. It is our experience that the statistical approach results in more distinct categories than the group consensus approach alone. The two approaches will treat disagreement differently. The statistical approach will tend to resolve sorter disagreements by creating new categories provided at least a few sorters (but not all) agree with one another, more or less. Disagreements in the group consensus are more usually resolved via some negotiation in the group setting. This negotiation usually results in fewer, not more, categories as well as the perspective of the "winning" negotiator or group facilitator.

Also, in our experience conducting sorts to define a variety of domains in a number of countries (from leadership behavior to parents' assessments to indicators of emotional health in the United States, South America, Africa and Asia), the results of this analytic process are sensible and readily accepted by the participants. While it can be computationally cumbersome, once presented to the group, it facilitates discussion and enables a quick group consensus on the results.

The statistical approach is even more democratic and participatory in the sense that all sorts and all incidents are weighted equally and simultaneously in the analysis; therefore, particulars of group dynamics (e.g., gender, personality characteristics, power, and other group and attention processes) do not bias the results. It provides structure to the discussion and consensus that, in some cases, would never happen otherwise.

As a mixed qualitative-quantitative method, it leverages strengths of both while, at the same time, has weaknesses in each area. It tends to be more systematic and rigorous that most purely qualitative methods; it tends to be more open-ended and flexible than most quantitative methods. For those of us in the field, the method has a lot of practical utility.

The rest of the document provides a procedure for creating the correlation matrix, conducting the analysis, and with some attention to sort activities. It offers step-by-step instructions and is oriented towards practical application. Periodically, the steps will include various subsections called *Note.* The notes provide additional information that is informative but not essential to performing the steps. The steps around the creation of the correlation matrix are more detailed as regular statistical software does not automatically create this matrix. Therefore, this phase involves more manual steps; however, we (the research team at the Center for Victims of Torture; research@cvt.org) have an Excel tool that will calculate this matrix. Once this matrix is created, it is a relatively straight-forward matter to import the matrix

into a statistical software package for subsequent data reduction using any number of conventional techniques. Materials on these procedures are widespread.

Sort Activities Review

1. It is largely assumed that the domain of interest has been specified and specific question(s) have been created (see Brief Ethnographic Interviewing Manual). Also, responses, or specific behavioral examples, are collected from members of the population of interest. We call these responses *incidents*. The method works the best if the given question engenders responses that are specific, behavioral (i.e., can be observed), and concrete. The method becomes less useful if incidents are in abstract terms, contains generalities, or are non-specific.

Note: If your goal is to explore some narrow facet or phenomena (e.g., how young children misbehave in school), it is best to ask for specific examples of children misbehaving school. If the goal is to broadly define some thing, it is better to ask broad questions (e.g., incidents about good parenting). Eventually, most diligent sorters given good incidents tend to create 10-20 categories to explain all the incidents. The more specific responses will result in more nuance and specificity in the categorical structure. Broader responses will engender broader categories. For example, at CVT in Minneapolis, MN, we asked for examples for how we know clients are doing well. One category was about having basic needs met, including winter clothing needs (Minnesota is very cold in the winter). This was sufficient for our purposes of broadly exploring domains of client functioning as a function of basic needs. However, it would be insufficient if we were primarily interested only in clothing needs. Except for identifying clothing as a part (of several) basic needs, it didn't provide much more detail. If we were only interested in clothing, we would ask specific question and possibly uncover clothing needs related to winter, clothing needs related to employment, the general condition of the clothing, the capability to purchase one's own clothes, clothes for religious services, etc. So, for understanding the global functioning of our clients, our approach worked very well. This was our goal. However, if we wanted to investigate clothing needs specifically, the results the approach would have been inadequate. One should carefully consider the scope of one's inquiry and make sure the question elicits responses reflective of that desired scope.

2. Incident reports are edited to be sensible (Bownas and Berdain 1988; Olson, 2000; Vinson, 2006). Such edits are usually minor but can include more substantive edits and selectively eliminating incidents. The most common edit is creating multiple incidents from one response. During much open-ended responding to questions, participants often provide more than one observable thing (Bownas & Berdain, 1988). Incident reports that contain multiple behaviors tend to be more haphazardly sorted across all sorters, resulting in more error in later analyses. Nonetheless, editing can introduce bias if taken too far. At the same time, incidents that describe people's personalities, describe general characteristics of people, provide insufficient detail, provide too much contextual information, include judgment, are fragmented, are non-sensical (check with members of the relevant

population), or just do not answer the question are not useful and can be edited or removed. It is a judgment call when/if to edit or remove an incident. Generally, less is better. Provided you do not have a preponderance of incomprehensible incidents, the analyses of sort data will also reveal if such incidents are equally incomprehensible to members of the population (or just the researcher).

Our experience has been that training incident providers on the method, providing a context for the effort, and collecting the incidents in-person tends to yield better responses.

Note: In general, a larger pool of incidents will uncover a greater variety of categories or is more likely to uncover less common occurrences. Therefore, one would think that more incidents are better. However, this is not the case. A good target for the number of incidents is around 200-250. Sorters can have difficulty, sort more haphazardly, or start to rebel if the number of incidents substantially exceeds this number. If you collect far more than 200 incidents, *randomly* select a subset of approximately 200-250 incidents for the sort. It is better to have 200 incidents diligently sorted than 300 where the sorting is more haphazard. This restriction dovetails with the earlier note to make sure the scope of your inquiry is restricted to what you want to know. Stated alternatively, you have 200 incidents to uncover as much as you can about some thing. If the question in your incident collection is too broad, you may not gather enough incidents to provide much nuance about that one thing.

3. Sorters perform independent sorts where they sort the incidents into homogenous categories of their own choosing. It is important that sorters complete these sorts independently and do not influence each other. If a group consensus process is also used, it should be used after the independent sorts. If at all possible, there should be at least 10 sorters. The number of sorters actually needed depends on the clarity of the underlying categorical structure and agreement between the sorters. Disagreement between sorters or sorts involving more complex/abstract phenomena are more readily resolved, in analysis, with more sorters. We have had sensibly interpretable results in all cases where I've used at least 10 sorters. As with most things statistical, more is safer. Each sorter should have his/her own set of incident cards to sort. At the very least, each sorter's cards should and can be stored as sorted (e.g., using rubber bands and envelopes). The subsequent analytic process is very picky about making sure all cards are represented in all the sorts. There will inevitable data entry errors related to the sort that must be corrected; storing each sorters' cards ensures that corrections can be made. A good sort should take a few hours and several iterations as sorters read through the incidents, come up with categories, revise, and resort.

Note: Picking sorters is another task that can affect results. In some cases, sorters may be obvious: members of the community. However, make sure your sorters reflect the population whose perspective you want. In some cases, this may mean members of the community, it could mean mental health professionals, it may be just women, men, etc. Make sure the sorters are motivated to diligently complete the task. Sorters should be literate (yes, it has happened). A well-done

sort usually takes a few hours and multiple iterations of sorting and re-sorting incidents. Sorters should be asked to define and describe each category they create. This can help you in defining categories later; however, it also provides some structure for the sorter to follow. A poorly conducted sort will tend to a) have fewer categories, b) be defined less well, and c) may have many incidents sorted into a miscellaneous category. Conducting sorts together (but working independently) can be a way to enhance sort quality, especially in places where solitary-style activities are less prevalent. However, care needs to be taken to gently redirect sorters away from influencing each others' sorts.

Distilling Sort Data into Categories Pt. I (Creating of Correlation/Agreement Matrix)

To distill categories from the sorts, the aforementioned correlation matrix between all sorted incidents is constructed for all the sort data. We will call this an *agreement matrix* for the rest of the document as the values in the matrix reflect the agreement between sorters. The following contains steps to construct the matrix, manually. It is worth noting that there is an Excel tool available that can construct the matrix (steps 5 through 7). It can be obtained from the research department at the Center for Victims of Torture (research@cvt.org). Nonetheless, it is worth reading through and understanding the following steps as an aid to understanding the method.

Note: Another reason we call this an agreement versus a correlation matrix is that the matrix will not exactly replicate the properties of a correlation matrix based on the most common type of correlation, Pearson's *r* between two continuous variables.

4. The first step is to create a count matrix for each sorter. For our example, let us suppose that we have 250 incidents sorted by 10 sorters. You would create 10 matrices, one for each sorter. Each matrix is a 250 x 250 matrix, where each incident (1-250) has a corresponding row and a corresponding column. The symmetrical matrix is created by entering a "1" into the matrix at points where two incidents were placed into the same category by a given sorter. A "0" is entered into the matrix at points corresponding to pairs of incidents that were placed into different categories. For example, let's suppose that a sorter places incidents 25 and 35 into the same category, but incident 45 is into a separate category. Then, the points of the matrix that correspond to rows/columns 25 and 35 would contain a "1". The points of the matrix that correspond to rows/columns 25 (and 35) and 45 would contain a "0." In this manner, a matrix should be constructed for each sorter.

5. The next step is to create a single, overall, sorter agreement matrix, called the *direct agreement* matrix. We will continue with the example of 250 incidents and 10 sorters. You simply add all of the matrices created in the first step, and divided each matrix value by 10 (i.e., the number of sorters). This results in one 250 x 250 matrix. Now, each point in the matrix is a proportion of how many of the sorters put two given incidents into the same category. For example, if 5 of 10 sorters placed incidents 25 and 35 into the same category, the value at the

intersection of rows/columns 25 and 35 would be .50, the number of judges (5) who put the incidents in the same category, divided by the total number of judges (10). If all judges put incidents 25 and 35 into the same category, the value would be 1.0. Similarly, if no judges put these incidents in the same category, the value would be 0, and so on.

The agreement matrix is akin to a correlation matrix between the incidents, where higher values represent higher levels of agreement for incident pairs among sorters. Lower values represent lower levels of agreement. However, unlike a correlation matrix, negative values are not possible. Individual values deviate from a true zero point (no agreement). In other words, a negative relationship between incidents would not make sense, because sorters cannot have a negative level of agreement. They either agree or they do not agree (Vey, 2003). A 0 indicates complete disagreement between all sorters.

6. You can subject the above matrix, the *direct agreement* matrix, to a data reduction method (see later steps). However, it is advantageous (for results and later analysis issues) to transform this direct agreement matrix into a 250 x 250 *standardized mean inner product (SMIP) agreement* matrix (Borman & Brush, 1993; Olson, 2000; Rod Rosse, Ph.D., personal communication, Winter 2006). We will explain the process here, as one could calculate manually. However, if you desire to do this, we highly recommend either contacting the research team at CVT (research@cvt.org) for the excel tool or employing someone adept at programming computers to perform the matrix algebra. Calculating this matrix literally involves tens of thousands of calculations.

The SMIP matrix is similar to the direct agreement matrix constructed in the previous step. There is one important difference. A given value at a point in the matrix no longer merely represents the proportion of judges that put those two incidents into the same category; it represents the pattern of similarity between the *first incident* to all other incidents in the matrix with the pattern of similarity between the *second incident* and all the other incidents in the matrix. (Olson, 2000). To use the earlier winter clothing example, it just doesn't consider how often a *scarf* incident and a *gloves* incident are paired together. It also considers how often both *scarf* and *gloves* are paired with each other and *coat, hat,* and *boots* as well. Here is a more detailed but more mundane example. Let us consider hypothetical incidents numbers 25 and 35. In the SMIP agreement matrix, the number at the intersection of rows/columns 25 and 35 now represents the degree to which incidents 25 and 35 were similarly categorized across all incidents. In other words, if sorters tended to categorize incident 25 with incidents 35,45,55,65,75, & 85, and sorters also tended to categorize incident 35 with incidents 25,45,55,65,75, & 85 (i.e., all the same incidents), the value at the intersection of rows/columns 25 and 35 would be relatively large. Alternatively, let us suppose that sorters categorized incident 25 with incidents 35, 45,54,64,74, & 84 and categorize incident 35 with incidents 25, 45,56,66,76, & 86 (i.e., half the same, half different). In other words, while incidents 25 and 35 tended to be categorized with each other and another incident, these incidents did not share as many incidents in common as in the previous scenario. The value at the

intersection of rows/columns 25 and 35 would be relatively smaller. In sum, the values in the SMIP agreement matrix summarize the relationships between incidents' ratings across all other incidents' ratings, not just the agreement for a given pair of incidents (Rod Rosse, Ph.D., personal communication, Winter 2006). This results in a symmetrical 250 x 250 matrix in which the diagonal contains values of one, while off-diagonals contain values ranging from zero to one, with higher values representing higher levels of agreement between judges (Vey, 2003). This matrix allows for more detail in the matrix.

The equation used to arrive at the standardized mean inner products is identical to the equation for the Pearson correlation coefficient except that the means are not subtracted from the individual elements. Means are not used as a reference point as the values have an absolute zero reference point (i.e, no agreement between all sorters).

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

If you will be calculating manually, inputting either the pure count matrix (e.g., 8 sorters put the incident in this category) or the proportion matrix (.80, for 8 out of 10 sorters) will yield the same results. We will assume the proportion value in this description. X and y each represent the two incidents where you want to calculate the "correlation", *r*. If you have the 200 incidents in our example, you will have to repeat the above equation for each possible incident pairing.

The example below goes through how you would calculate this for one pair of incidents. Let's assume you are trying to calculate the *r*-value between incident #1 (x) and incident #2 (y). In the numerator, the sum of the product between incident 1 and all other incidents, and incident 2 and all other incidents. So, if incident 1 has a proportion value of .50 with incident 4 (i.e., 5 of 10 raters put incidents 1 and 4 in the same category) and incident 2 has a proportion value of .40 (4 of 10), this value would be .25. Like values are calculated for all the relationships that incidents 1 and 2 have with all other incidents. This is your numerator value for the right-side of the equation, between incidents 1 and 2. The denominator uses the same values as in the numerator; however, they are squared and summed independently, before being multiplied, and then raised to .5 power (i.e., square root).

It isn't a bad exercise to make a very small matrix (e.g., 5 or 6 incidents) to manually work through the math and understand what is going on. However, for real data, use a computer.

Note: The argument for preferring the SMIP agreement matrix as opposed to the direct agreement matrix is that it better captures the relationships between all the items and captures somewhat more nuance in the relationships between the incidents in the sort. For each point in the matrix, it also allows for a term with

more possible discrete values. For example, a direct agreement matrix with 10 sorters allows for 11 possible values for each pair (0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1.0); the SMIP allows for thousands of discrete values. In practice, using the SMIP matrix does enhance the fit of the eventual data reduction model. However, for more robust categories (i.e., the most apparent or strongly agree-upon categories with many incidents), the results tend to be similar between the SMIP and the direct agreement matrices.

Distilling Sort Data into Categories Pt. II (Analyzing Correlation/Agreement Matrix))

7. The next step is to determine data-driven categories from the sort data using the agreement matrix from the previous step. Because this matrix is akin to a correlation matrix, common data reduction methods include methods such as factor analysis, principle components analysis, or multidimensional scaling. With factor analysis, only the variability that items have in common with other items are used. Factor analysis reduces data according to the proportion of variance of a particular item that is due to common factors (i.e., common variance). This type of data reduction is used when one postulates that relationships between numerous observed variables are due to some smaller set of latent, underlying, unmeasured variables. Alternatively, with principle components analysis, it is assumed that *all* variability in an item should be used in the analysis (i.e., total variance) to derive a set of components that explains the greatest proportion of that total variance. Because the goal of our effort is to uncover categories in the data with the goal of explaining the most total variance, principle components analysis is preferred. However, a cogent argument can be made for factor analysis if that is your preference. Practically, a factor analysis is computationally more complex and can introduce some additional problems (see part 8.c.). Practically, results between the two techniques are often quite similar. In some ways, multidimensional scaling may be the preferred method as it is better oriented towards handling a matrix with non-negative values (like our matrix) with fewer issues. However, principle components analysis is more likely to be understood by most audiences; information about this technique is more commonly available in most multivariate statistics books. However, if you have a familiar analyst, we would also try a multidimensional scaling technique in the analysis.

Note. When you input the matrix into a statistical package, use the number if incidents as the N for the sample. Determining the appropriate N, with this type of analysis, is not as straight-forward as when we have a regular sample with some measure for X number of people. Specifying a larger N will tend to be advantageous in estimating a model. However, this means you should not trust various fit statistics that your software may produce with the model. For your sample of incidents and the categorical structure, judge how well the model fits by the theoretical sensibility and utility of structure as compared to other models with varying numbers of factors/components.

8. Once you determine a method, here, assumed to be principle components analysis, you have all the options of that form of analysis. We would like to cover a few of the main considerations: rotation, number of components/categories, and

the failure of a model to converge or be positive definite. Alternative materials on these analytic techniques are widely available in just about any multivariate analysis reference.

8.a. Rotation. In principle components analysis, rotation concerns the algorithm used to determine a final structure solution for the data. A discussion of rotations is beyond the scope of this chapter. Nonetheless, there are two basic types of rotations, orthogonal and oblique. Unless you are an experienced user or seek additional information to determine which type of rotation to use, we suggest using the orthogonal rotation, varimax. This is appropriate given that we are working to make either/or categorical decisions for all incidents. It is also the predominant rotation used in this type of analysis and the default option in most statistical packages.

8.b. Number of components/categories. Components are the statistical analog to the categories we will derive from the analysis. The analyst has to decide how many components or categories he/she wants. This is the same process one undertakes with factor analysis. There are various rules concerning the number of components/factors such as the so-called Scree Test or Kaiser Criterion. These rules often make the choice seem more of an empirical issue when it is actually an empirically-informed judgment. The analyst usually picks one amongst several models that, more or less, fit the data similarly but where one model makes more sense given some theoretical context. For this type of analysis, the general rules we use for adopting a specific principle components solution in this context are: 1) each component in the solution has to have at least one incident that loads primarily on that component, 2) incidents loading on each component have to be similar to one another, 3) each component has to account for more variance than a single item (i.e., the Kaiser criterion), and 4) it makes sense given the context.

Note. You will have to iteratively run many models to determine the number of components you want. We suggest starting with a higher number (e.g., 20), then pairing it back to fewer components that make sense within the guidelines provided above. When working with a group, we usually present a final set of 3-4 final models (with some definitional interpretation) and work with the group to select the final, accepted, model.

8.c. Model fails to converge or is not "positive definite". Earlier, we commented that the matrix we create from this analysis is not a true correlation matrix in the sense of Pearson's correlations between continuous variables; however, this analysis treats it like it is. Such issues are not unheard of, especially with data using polychoric or tetrachoric correlations (common in clinical research). We can certainly observe them with sort data. However, we do not want to alarm you, this may easily never be encountered. Also, convergence issues are less likely to be an issue if you use principal components analysis versus a factor analysis. Using the SMIP matrix should also help in most situations. If using factor analysis, avoid Generalized Least Squares (GLS) of Maximum Likelihood (ML or MWL) extraction methods in favor of Principal Axis Factoring (PAF). A Multidimensional Scaling technique may also sidestep the issue. Making some changes (i.e., making

larger) the input N for the analysis may also help. Lastly, different software packages can respond differently to these issues altogether, from allowing you to override, simply providing a warning, or ceasing to work altogether. Try using a different statistical package. If it continues to be an issue, you may want to seek a statistical consult, double-check the data, and/or do some further reading into your data reduction technique.

Note. Some of this may make you queasy. It seems analytically reckless to consider tweaking analyses, changing your input N, or just shopping around to different reduction methods or statistical packages just to get a model to converge. In research, in particular, it is easy to come up to believe there is the right way to statistically analyze the data. Especially for factor analytic and related methods, we can start to see this as uncovering some underlying, latent, truth. Deviations from this seemingly undermine the truth of what we are finding. An alternative conceptualization is that these are simply different ways of summarizing patterns in numbers. Are those summaries useful? The statistical adage that all models are wrong but some models are useful can be applied here. The primary concern is getting a sensible reduction of the interview data and getting results that are incrementally better and of greater utility than the alternative. Here, the alternative could be the group consensus approach to the sort data. For example, is forcing a positive covariance from a slightly negative one better or worse than a group consensus session largely dominated by a cadre of three, strong-willed, extraverted males? As with any analysis, there are judgment calls, both explicit and commonly, implicit. Here, they can become very apparent. This apparentness is one of the occupational hazards of engaging thoughtfully in a mixed qualitative-quantitative method. Again, none of this may ever come up. Nonetheless, you could find yourself in the situation where you are tweaking some aspects of the data and are sensibly worried that they could be affecting the results in some substantive, unknown, way. If so, a statistical consult could ease concerns, redress issues specific to your data, or provide some other method to analyze the data.

Note: When we perform this step with a population of which we are not a member, we will produce 3-4 models that, from the data, we think are the top contenders. Then, we will present and interpret these models for our sorters or other colleagues from that population. Together, we will select the final model. Also, if you use conventional statistical software, it will take some effort to compile, print, and create a single table that contains all factor loadings and item information in a way that allows for interpretation.

9. After a final model is selected, categories should be defined using the constituent incidents as well as integrating the category definitions provided by individual sorters. It is also useful to provide some examples of each category from the descriptions provided. Be very careful about confidentiality issues of provided examples if this is an issue for your community of incident providers.

10. A rarely exercised *optional* step that can be considered if the derivation of the categorical structure, itself, is of primary interest is a *retranslation* sort. All of step 10 considers this optional step.

You have another set of independent (uninvolved to this point) sorters re-sort all the incidents into the categories defined in the previous step. The goal is to reliably classify all the incidents back into the categorical structure as a check and "test" of that structure. Moreso, you use it to refine and the categorical definitions created in step 9. The retranslation sort requires fewer sorters than the initial sort. Moreover, it is feasible to "re-sort" more incidents into this categorical structure than used in the initial sort (if you had to randomly select some subset of 200 incidents from a larger set of incidents). You will have to devise some decision rule that specifies what level of agreement, between translation re-sorters, you require before considering an incident reliably re-sorted without further discussion. This further discussion occurs in a consensus meeting between all retranslation sorters and the analyst/researcher. This meeting should occur after all retranslation sorters have completed their sorts, agreement has been determined, and the research/analyst has had the time to prepare materials specifying difficult incidents for discussion at the meeting. Through this process, you tend to both facilitate agreement of difficult incidents as well as refine the definitions of your categories.

For example, let's say we have 5 retranslation sorters. If at least 4 of the sorters agree that a given incident belongs in a certain category, it is considered reliably sorted. However, if less than 4 agree, then that incident has to be discussed at a consensus meeting between retranslation sorters. The amount of required agreement (e.g., 4/5) is a suggestion; your rule or number of sorters can vary. In a good sort with good subsequent categorical definitions, the vast majority of incidents should be reliably resorted. You should be discussing, at most, between 10%-15% of the incidents, preferably far less. If you find substantially less agreement, consider if there were issues in data collection, the initial sort, analysis, category definition creation, or something with the retranslation sorters. At the meeting, the sorters have to come to some agreement as to its appropriate place. Sometimes, the definitions created need some editing or revision; this is the most beneficial part of the retranslation process. Sometimes, incidents are too ambiguous or contain content reflective of multiple categories and cannot be reliably sorted (FYI, these are most likely to be the items with low communalities from the data analysis).

Note: As an activity involving agreement between multiple raters, a number of statistical measures of agreement or reliability are available aside from simple percentage of sorters that agree (e.g., Cohen's kappa, Krippendorf's alpha, etc.). We have also found the use of such statistics useful when determining if we have certain categories that seem to be sorted at higher levels of agreement. However, these statistics are likely the most optional part of this optional step.

11. Congratulations. You are done with your sort and have your categorical structure.

Note: There are other analytic options beyond creating this set of categories, including the creation of a hierarchical categorical structure. In other words, we can empirically nest the categories into larger, superordinate, categories using some

type of hierarchical analysis with the sort data. This is beyond the scope of this manual; however, the various data reduction methods depicted here do allow for this possibility. This may be worth considering if the primary purpose of your effort is to define relevant domains and/or further reduce an expansive categorical structure to something with fewer categories. However, if the primary purpose is measure development, we encourage you select the most relevant categories from the primary analysis to develop your initial (sub)scales. For measurement purposes, superordinate categories tend to result in overly abstracted constructs that tend result in more muddled scores with decreased reliability. Our observation is that items will still tend to empirically differentiate themselves in accordance with the primary categorical structure, to the detriment of the higher-level scores (unless you have many items).

Note: If you have questions or would like additional support, please contact the research team at the Center for Victims of Torture at: research@cvt.org or www.cvt.org

References

Anderson, L. & Wilson, S. (1997). Critical incident technique. In D. Whetzel & G.R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 89-112). Palo Alto, CA: Davies Black Publishing.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6*(1), 1-21.

Bownas, D. A. & Berdain, H. J. (1988). Critical incident technique. In G. Sidney (Ed.), *The job analysis handbook for business, industry, and government* (pp. 1120-1137). John Wiley & Sons, Inc.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51(4),* 327-358.

Olson, A. (2000). *A theory and taxonomy of individual team member performance*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Vey, M. (2003). *Motives and rewards of effective contextual and task performance.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.